Full Title: Evolution is in the details: Regulatory differences in modern human and Neanderthal

Short title: Regulatory differences in modern human and Neanderthal

Harlan R. Barker [1,2*], Seppo Parkkila [1,2], Martti E.E. Tolvanen [3]

1 Faculty of Medicine and Health Technology, Tampere University, 33520 Tampere, Finland.

2 Fimlab Ltd., Tampere University Hospital, 33520 Tampere, Finland.

3 Department of Computing, University of Turku, 20014 Turku, Finland.

*To whom correspondence should be addressed.

Abstract

Transcription factor (TF) proteins play a critical role in the regulation of eukaryote gene expression by sequence-specific binding to genomic locations known as transcription factor binding sites.

Here we present the TFBSFootprinter tool which has been created to combine transcription-relevant data from six large empirical datasets: Ensembl, JASPAR, FANTOM5, ENCODE, GTEX, and GTRD to more accurately predict functional sites. A complete analysis integrating all experimental datasets can be performed on genes in the human genome, and a limited analysis can be done on a total of 125 vertebrate species.

As a use-case, we have used TFBSFootprinter to study sites of genomic variation between modern human and Neanderthal promoters. We found significant differences in binding affinity for 110 transcription factors, which are enriched for homeobox and brain. Analysis of single cell data show that a subset of these (CUX1, CUX2, ESRRG, FOXP1, FOXP2, MEF2C, POU6F2, PRRX1 and RORA) co-occur as marker genes in L4 glutamatergic neurons.

Differential binding sites for these transcription factors were found in 74 target genes, the largest number of which were found in the bidirectional promoter of key mitochondrial-function genes FARS2 and LYRM4.

Abbreviations

Assay for Transposase-Accessible Chromatin using sequencing (ATAC-Seq)

Cap analysis of gene expression (CAGE)

Chromatin immunoprecipitation with massively parallel DNA sequencing (ChIP-Seq)

Cis-regulatory elements (CREs)

Database of Immune Cell eQTLs (DICE)

High-throughput systematic evolution of ligands by exponential enrichment (HT-SELEX)

Position weight matrix (PWM)

Transcription factor (TF)

Transcription factor binding site (TFBS)

Transcription start sites (TSSs)

Weeks post conception (PCW)

Introduction

North of Eden

Humans and their hominin relatives have been leaving Africa in waves for the past two million plus years. Various environmental and cultural pressures have impacted each diaspora in different ways, subsequently producing adaptations reflected in physiology, immunity, and brain size. In relatively recent history, the discovery and sequencing of DNA from remains of Neanderthal [1–4] and Denisovan [5,6] has now allowed direct comparison of DNA from modern and ancient hominids. In the observed genomic variations between modern humans and Neanderthals, a limited number have been identified which occur in gene coding regions. Some of these are found in genes known to affect cognition and morphology (cranium, rib, dentition, shoulder joint) [1], pigmentation and behavioral traits [2], and brain development [7]. However, as has been noted before, there is a paucity of coding variations to explain the differences between related species; the genome of the Altai Neanderthal reveals just 96 fixed amino acid substitutions, occurring in 87 proteins [7]. Unsurprisingly, a much larger set of variants are observed in intergenic regions, owing not only to the fact that these are comparatively much larger regions but also to the expectation of lesser conservation in what until recently was often termed "junk DNA". While variants in coding regions can directly affect protein structure, those found in intergenic regions may affect regulation of gene expression, through alternative binding of transcription factors in promoters and enhancers and expression of non-coding RNAs. What may surprise some, is the cumulative effect of numerous small — and large — changes to gene expression arising due to these manifold intergenic changes, and which may ultimately serve as the engine of speciation. Indeed, analysis of introgressed Neanderthal DNA is more depleted in regulatory regions than in those which code for proteins [8]. Using a new computational tool we have created, which incorporates numerous

transcription-relevant genomic features, we sought to reveal how the comparative differences in these regions may affect regulatory differences between modern humans and Neanderthals. Specifically, our aim is to identify those gene-regulating transcription factors whose binding to DNA may vary between these species of hominid and thus drive the differences between them.

Transcription factors drive gene regulation

As early as 1963, Zuckerkandl and Pauling began addressing the apparent disparity in the fact that species with obvious differences could have proteins which look so similar. At the time they posited that this could be explained by the idea that "in some species certain products of structural genes inhibit certain sets of other genes for a longer time during development than in other species." [9]. Since this early exposition, the idea was more formally proposed in a publication by King and Wilson [10], and now the importance of regulatory regions in evolutionary adaptation has been further explored and accepted [11,12].

Transcription factor (TF) proteins play a critical role in the regulation of eukaryote gene expression by sequence-specific binding to genomic locations known as cis-regulatory elements (CREs) or, more simply, transcription factor binding sites (TFBSs). TFBSs can be found both proximal and distal to gene transcription start sites (TSSs), and multiple TFs often bind cooperatively towards promotion or inhibition of gene transcription, in what is known as a cis-regulatory module (CRM). Because of the role these proteins play in transcription, discovery of TFBSs greatly furthers understanding of many, if not all, biological processes. As a result, many tools have been created to identify TFBSs. Owing to the time and material requirements of individual mutation studies needed for experimental verification, many of these tools are computational. However, at issue in both the synthetic and experimental approach are two distinct problems. First is identification of where TF proteins bind to DNA. Experimental tools like chromatin immunoprecipitation with

massively parallel DNA sequencing (ChIP-Seq) are rapidly revealing the landscape of TFBSs for individual TFs in various cell types, and under various conditions. Computational tools often leverage this new data to build TFBS models which are increasingly accurate at making those predictions in silico. The second issue is how to determine if an experimental site (e.g., ChIP-Seq peak) or putative computational prediction (statistically likely) are actually biologically relevant, such as contributing to gene expression/repression, chromatin conformation, or otherwise. This second problem is by far the more demanding of the two, and it is here where we seek to levy the inclusion of transcription-relevant data.

The law of large numbers

Computational prediction of TFBSs seeks to enlist experimental data in the quest for the best possible specificity and sensitivity. However, computational modeling has deficits rooted in the law of large numbers; that is, because of the large size of a target genome, any method of prediction is bound to produce a large number of spurious results and filtering or thresholding results often means lost true positives. This can be compounded by the fact that biologically relevant TFBSs can be weakly binding [13].

Depending on the approach, the extent of incorporation of relevant experimental data varies widely. From very early on, the position weight matrix (PWM) has been used to represent and predict the binding of proteins to DNA. PWMs use a single, but very relevant, type of experimental data, that derived from observed binding events [14]. To create a PWM, a count is made of each of the four nucleotides at each position in the set of experimentally determined binding sites, known as a position frequency matrix (PFM). With the PFM, and using some contextual information about the target genome, a PWM probability model is generated which represents the binding preferences of a TF (detailed in Methods). The PWM can then be used to arrive at a likelihood

score for a target DNA region, which thus represents the likelihood of a TF binding to that DNA sequence. The accuracy of this method can continue to improve solely due to the large, and increasing, amounts of TFBS sequencing data provided by newer experimental technologies; like ChIP-Seq [15], high-throughput systematic evolution of ligands by exponential enrichment (HT-SELEX) [16,17], and protein binding microrarrays (PBMs) [18].

Old game, new tricks

Updates to the traditional PWM have appeared over the years. One addresses the fact that traditional PWMs assume that the binding preference of a TF is independent at each position of the sequence it binds, resulting in both dinucleotide [19,20] and position flexible [21–23] models. These become relevant when a TF protein has significantly different binding modes due to changes in conformation, structure, or splicing [24]. Ultimately, however, the difference between position independent and dependent models has been shown to be minimal, and TFs with more complex binding specificities can be accounted for using multiple position-independent PWMs [25,26].

New statistical and computational algorithms, as they have come in vogue, have been brought to bear on the problem of TF-DNA binding: regression, Monte Carlo simulations, Markov models, machine learning, and deep learning, to name the most prominent. An important distinction is that these, and many other, approaches seek to improve what their creators viewed as the primary aspect of prediction of TFBSs, the binding model itself. There has been great success in targeting this aspect, and while it is critical to know, the location of binding alone is not indicative of function [27]. Because of the incremental gains which have been achieved with new computational modeling of TF-DNA binding events, it can be argued that further work in this space is best served in discovery of binding sites for TFs which are not currently cataloged or modeled. Indeed, this is often a key difference to be noted when choosing between traditional PWMs and some of the

later computational tools/models. The existence of a greater number of TF binding models is a strong reason to choose to incorporate the more extensively used traditional PWMs in binding prediction.

In the search for increased accuracy, other new models have improved TFBS prediction by instead incorporating other relevant biological data, for example: 3D structure of DNA [28–32], chromatin accessibility/DNAse hypersensitivity sites [33,34], overlap in gene ontology [35], amino acid physicochemical properties [36], and gene expression and chromatin accessibility [37,38]. These alternative models often match or outperform strictly sequence-based models [23,32].

TFBSFootprinter incorporates transcription-relevant data

We sought to identify multiple sources of experimental data relevant to gene expression and transcription factor binding, and to incorporate it into a comprehensive model in order to improve prediction of functional TFBSs. Specifically, clustering of TFBSs has been shown to be an indicator of functionality [27,39,40]; conservation of genetic sequence across genomes of related species is one of the most successfully used attributes in identification of TFBSs [40,41]; proximity to TSS is strongly linked to TFBS functionality [42]; correlation of expression between a transcription factor and another gene is an indication of a functional relationship [37,43,44]; variants in non-coding regions have a demonstrated effect on gene expression [45–47] and variants affecting gene expression are enriched in TFBSs [48]; open chromatin regions (ascertained by ATAC-Seq or DNAse-sensitivity) correlate with TF binding [49]; and finally, as previously mentioned, significant effort has gone into identifying the actual composition of the binding sites themselves through the use of sequencing of TFBSs (e.g., ChIP-Seq and HT-SELEX) [16,22,50].

Ensembl identifier-oriented system of analyses

For our tool, instead of using simple absolute genomic coordinates, the Ensembl transcript ID was chosen as the basic unit of reference. This is useful for several reasons. First, the Ensembl database is one of the most well-maintained biological databases in existence. It is continually updated and expanded and contains a wealth of sequence and regulatory information on a large number of vertebrate species. As a result, the tool we present here — TFBSFootprinter — can offer predictions in 125 vertebrates at the time of writing. From human and model organisms such as mouse and zebrafish, to African bush elephant, the catalog will increase as the Ensembl database itself expands. Second, it allows the inclusion of important datasets which are gene-centric, such as FANTOM TSSs and expression data, GTEx eQTLs, and all annotations which are compiled within Ensembl itself. Finally, the Ensembl transcript ID provides an easy point of reference for a greater audience of scientists, thus increasing the accessibility and utility of the tool.

Non sum qualis eram

This study supports the hypothesis that future advances in prediction offered by incorporation of transcription-related biological data will outshine that which can be achieved by improvements in modeling of TF-DNA binding alone. We show here that the TFBSFootprinter tool provides a good way to predict TFBSs based on incorporation of a variety of relevant biological data. As a proof of usage, we apply TFBSFootprinter in a comparative analysis of locations of variation in the promoters of modern humans and Neanderthal genomes.

Results

High-scoring TFBSs differ between modern human and Neanderthals and are enriched for homeobox and brain

In analysis of 21,990 SNPs occurring in comparison of the modern human and Neanderthal promoteromes — the collection of all human/Neanderthal proximal promoters of protein-coding genes — a total of 108 TF models, representing 110 unique TF proteins, showed a significant difference in scoring between the two human species (Table 1). Compared to all protein-coding genes, the complete set of TFs cataloged in the JASPAR database are very strongly enriched for homeobox genes (Fisher's exact test: odds ratio [OR] 85.16; p-value 5.79 x 10^{-177}) and moderately for brain, eye, retina genes (OR 1.80; p-value 3.50 x 10^{-11}). Likewise, compared to the complete set of JASPAR TFs, homeobox genes are strongly over-represented among the differentially binding (DB) TFs (78/110; OR 12.40; p-value 2.96 x 10^{-17}). Similarly, genes with brain-specific expression (79/110) are also enriched in DB TFs (OR 2.86, p-value 3.48 x 10^{-6}) while JASPAR TFs themselves are enriched compared to all protein-coding genes (OR 1.80; p-value 3.50 x 10^{-11}). Immune-specific genes (29/110) were under-represented in DB TFs compared to the JASPAR set (Fisher's exact test: OR 0.37, p-value 1.44 x 10^{-5}).

Table 1. TFs with differential binding affinity in modern human vs. Neanderthal proximal and semi-distal promoter regions. FHBOX, forkhead box; HAR, human accelerated region; HBOX, homeobox.

Gene Symbol	Description	HBOX	FHBOX	Assoc. disease	SFARI Autism	GWAS Schizo.	HAR
ALX3	ALX Homeobox 3	x		Frontonasal Dysplasia 1*			
ALX4	ALX Homeobox 4	x		Frontonasal Dysplasia 2*			x

ARID5A	AT-Rich Interaction Domain 5A		Brachydactyly, Type E2			
BARHL1	BarH Like Homeobox 1	x	Triple-Receptor Negative Breast Cancer			
BARHL2	BarH Like Homeobox 2	x	Helicobacter Pylori Infection		x	x
BARX1	BARX Homeobox 1	x	Hypertrophic Pyloric Stenosis			
BSX	Brain Specific Homeobox	x	Jacobsen Syndrome			
CRX	Cone-Rod Homeobox	x	Cone-Rod Dystrophy 2*			
CUX1	Cut Like Homeobox 1	x	Global Developmental Delay With Or Without Impaired Intellectual Development*	2	x	x
CUX2	Cut Like Homeobox 2	x	Developmental And Epileptic Encephalopathy 67*	S, 2		
DLX1	Distal-Less Homeobox 1	x	Split-Hand/Foot Malformation 5			
DLX2	Distal-Less Homeobox 2	x	Axenfeld-Rieger Syndrome	2		x
DLX4	Distal-Less Homeobox 4	x	Orofacial Cleft 15*			
DMBX1	Diencephalon/Mesencephalon Homeobox 1	x	Farsightedness			
DMRT3	Doublesex And Mab-3 Related Transcription Factor 3		46,Xy Partial Gonadal Dysgenesis*			x
EMX1	Empty Spiracles Homeobox 1	x	Kallmann Syndrome		x	
EN1	Engrailed Homeobox 1	x	Endove Syndrome, Limb-Brain Type*			x
ESRRG	Estrogen Related Receptor Gamma		Usher Syndrome, Type Iia			x
ESX1	ESX Homeobox 1	x	Isolated Growth Hormone Deficiency Type Iii			
EVX1	Even-Skipped Homeobox 1	x	Hand-Foot-Genital Syndrome			
EVX2	Even-Skipped Homeobox 2	x	Synpolydactyly			x
FOS	Fos Proto-Oncogene, AP-1 Transcription Factor		Congenital Generalized Lipodystrophy*			

	Subunit						
FOXA1	Forkhead Box A1		x	Estrogen-Receptor Negative Breast Cancer			x
FOXD3	Forkhead Box D3		x	Autoimmune Disease 1*			
FOXK1	Forkhead Box K1		x	Colorectal Cancer			
FOXK2	Forkhead Box K2		x	Lung Cancer			
FOXP1	Forkhead Box P1		x	Mental Retardation With Language Impairment And With Or Without Autistic Features*	1	x	x
FOXP2	Forkhead Box P2		x	Childhood Apraxia Of Speech*	1	x	x
GBX2	Gastrulation Brain Homeobox 2	x		Colon Small Cell Carcinoma			x
GFI1	Growth Factor Independent 1 Transcriptional Repressor			Neutropenia, Nonimmune Chronic Idiopathic, Of Adults*			
GRHL2	Grainyhead Like Transcription Factor 2			Ectodermal Dysplasia/Short Stature Syndrome*		x	
GSC	Goosecoid Homeobox	x		Short Stature, Auditory Canal Atresia, Mandibular Hypoplasia, And Skeletal Abnormalities*			
GSC2	Goosecoid Homeobox 2	x		Digeorge Syndrome			
GSX2	GS Homeobox 2	x		Diencephalic-Mesencephalic Junction Dysplasia Syndrome 2*		x	
HAND1	Heart And Neural Crest Derivatives Expressed 1			Hypoplastic Left Heart Syndrome			
HMX1	H6 Family Homeobox 1	x		Oculoauricular Syndrome*			
HMX2	H6 Family Homeobox 2	x		Oculoauricular Syndrome			
HMX3	H6 Family Homeobox 3	x		Superior Semicircular Canal Dehiscence			
HNF1A	HNF1 Homeobox A	x		Maturity-Onset Diabetes Of The Young, Type 3*			
HNF1B	HNF1 Homeobox B	x		Renal Cysts And Diabetes			

			Syndrome*			
HOXA5	Homeobox A5	x	Embryonal Carcinoma			x
HOXB2	Homeobox B2	x	Omphalocele			
HOXB3	Homeobox B3	x	Breast Cancer			
HOXB5	Homeobox B5	x	Mowat-Wilson Syndrome			
HOXD8	Homeobox D8	x	Breast Cancer			
IRF3	Interferon Regulatory Factor 3		Encephalopathy, Acute, Infection-Induced 7*		x	
ISL2	ISL LIM Homeobox 2	x	Amme Complex			
ISX	Intestine Specific Homeobox	x	Hypervitaminosis A			x
JUN	Jun Proto-Oncogene, AP-1 Transcription Factor Subunit		Sarcoma			
LBX2	Ladybird Homeobox 2	x	Rela Fusion-Positive Ependymoma			
LEF1	Lymphoid Enhancer Binding Factor 1		Ectrodactyly And Ectodermal Dysplasia Without Cleft Lip/Palate*			
LHX3	LIM Homeobox 3	x	Pituitary Hormone Deficiency, Combined, 3*			
LHX4	LIM Homeobox 4	x	Pituitary Hormone Deficiency, Combined, 4*			
LHX9	LIM Homeobox 9	x	Gonadal Agenesis			
LMX1A	LIM Homeobox Transcription Factor 1 Alpha	x	Deafness, Autosomal Dominant 7*			
LMX1B	LIM Homeobox Transcription Factor 1 Beta	x	Nail-Patella Syndrome*	2		
MAFB	MAF BZIP Transcription Factor B		Multicentric Carpotarsal Osteolysis Syndrome*			x
MAFG	MAF BZIP Transcription Factor G		Homocystinuria-Megaloblastic Anemia Cble Type			
MEF2C	Myocyte Enhancer Factor		Neurodevelopmental Disorder	1	x	x

	2C		With Hypotonia, Stereotypic Hand Movements, And Impaired Language*		
MEOX1	Mesenchyme Homeobox 1	x	Klippel-Feil Syndrome 2, Autosomal Recessive*		
MEOX2	Mesenchyme Homeobox 2	x	Female Stress Incontinence		x
MIXL1	Mix Paired-Like Homeobox	x	Cardiofaciocutaneous Syndrome 2		
MNX1	Motor Neuron And Pancreas Homeobox 1	x	Currarino Syndrome*		
NEUROG1	Neurogenin 1		Cecum Carcinoma		
NFYA	Nuclear Transcription Factor Y Subunit Alpha		Renal Cell Carcinoma, Nonpapillary		
NKX2-5	NK2 Homeobox 5	x	Atrial Septal Defect 7 With Or Without Atrioventricular Conduction Defects*		
NKX2-8	NK2 Homeobox 8	x	Bronchus Adenoma		
NKX6-2	NK6 Homeobox 2	x	Spastic Ataxia 8, Autosomal Recessive, With Hypomyelinating Leukodystrophy*		
NOBOX	NOBOX Oogenesis Homeobox	x	Premature Ovarian Failure 5*		
NR1H3	Nuclear Receptor Subfamily 1 Group H Member 3		Multiple Sclerosis*	x	
NR2E1	Nuclear Receptor Subfamily 2 Group E Member 1		Enhanced S-Cone Syndrome		
NRL	Neural Retina Leucine Zipper		Retinitis Pigmentosa 27*		
ONECUT1	One Cut Homeobox 1	x	Pharyngoconjunctival Fever		x
ONECUT3	One Cut Homeobox 3	x	Breast Cancer		
OTX1	Orthodenticle Homeobox 1	x	Agnathia-Otocephaly Complex 2		
OTX2	Orthodenticle Homeobox	x	Microphthalmia, Syndromic 5*		

	2					
PAX3	Paired Box 3	x	Waardenburg Syndrome, Type 1*			x
PAX7	Paired Box 7	x	Myopathy, Congenital, Progressive, With Scoliosis*			
PHOX2A	Paired Like Homeobox 2A	x	Fibrosis Of Extraocular Muscles, Congenital, 2*			
PHOX2B	Paired Like Homeobox 2B	x	Central Hypoventilation Syndrome, Congenital, 1*			x
PITX3	Paired Like Homeodomain 3	x	Cataract 11, Multiple Types*			
POU1F1	POU Class 1 Homeobox 1	x	Pituitary Hormone Deficiency, Combined Or Isolated, 1*			x
POU2F1	POU Class 2 Homeobox 1	x	Herpes Simplex			
POU2F3	POU Class 2 Homeobox 3	x	Coronary Thrombosis			
POU3F1	POU Class 3 Homeobox 1	x	Lymphocele			
POU3F2	POU Class 3 Homeobox 2	x	Microphthalmia	S	x	x
POU3F3	POU Class 3 Homeobox 3	x	Snijders Blok-Fisher Syndrome*			x
POU4F1	POU Class 4 Homeobox 1	x	Ataxia, Intention Tremor, And Hypotonia Syndrome, Childhood-Onset*		x	
POU4F2	POU Class 4 Homeobox 2	x	Differentiating Neuroblastoma			
POU4F3	POU Class 4 Homeobox 3	x	Deafness, Autosomal Dominant 15*			x
POU6F1	POU Class 6 Homeobox 1	x	Clear Cell Adenocarcinoma Of The Ovary			
POU6F2	POU Class 6 Homeobox 2	x	Wilms Tumor 5*		x	x
PROP1	PROP Paired-Like Homeobox 1	x	Pituitary Hormone Deficiency, Combined, 2*			
PRRX1	Paired Related Homeobox 1	x	Agnathia-Otocephaly Complex*			x
PRRX2	Paired Related Homeobox 2	x	Hypoplastic Right Heart Syndrome			

RARA	Retinoic Acid Receptor Alpha		Acute Promyelocytic Leukemia*			
RAX	Retina And Anterior Neural Fold Homeobox	x	Microphthalmia, Isolated 3*			
RAX2	Retina And Anterior Neural Fold Homeobox 2	x	Cone-Rod Dystrophy 11*			
RORA	RAR Related Orphan Receptor A		Intellectual Developmental Disorder With Or Without Epilepsy Or Cerebellar Ataxia*	S	x	x
RXRA	Retinoid X Receptor Alpha		Hepatocellular Carcinoma			
SHOX	Short Stature Homeobox	x	Leri-Weill Dyschondrosteosis*	2		
SOX10	SRY-Box Transcription Factor 10		Peripheral Demyelinating Neuropathy, Central Dysmyelination, Waardenburg Syndrome, and Hirschsprung Disease*			
SOX15	SRY-Box Transcription Factor 15		Autosomal Dominant Non-Syndromic Intellectual Disability			
TBP	TATA-Box Binding		Spinocerebellar Ataxia 17*			
TCF3	Transcription Factor 3		Agammaglobulinemia 8, Autosomal Dominant*			
UNCX	UNC Homeobox	x	Spondylocarpotarsal Synostosis Syndrome			x
VAX1	Ventral Anterior Homeobox 1	x	Microphthalmia, Syndromic 11*			
VAX2	Ventral Anterior Homeobox 2	x	Microphthalmia			
VSX1	Visual System Homeobox 1	x	Corneal Dystrophy, Posterior Polymorphous, 1*			
VSX2	Visual System Homeobox 2	x	Microphthalmia, Isolated 2*			

DB TF target genes

A total of 74 genes were identified, hereafter referred to as ‘DB TF target genes’, whose promoters contain SNPs where differential binding occurs by our determined set of statistically significant DB

TFs (Table 2). The target gene with the greatest number of DB events (143) was FARS2, whose promoter contains 2 SNPs over which 105 (97.22%) of the DB TF binding models showed differential binding (Table 2). The remaining top ten DB TF target genes with the next highest number of DB occurrences were: TEX19, NARF, SLC6A11, LSM5, FAM172A, WNT7A, EXO1, KRT8, and ZNF37A. Of these, all but LSM5 have been identified with brain (FARS2, SLC6A11, FAM172A, WNT7A, and ZNF37A) or reproductive (FARS2, TEX19, NARF, WNT7A, EXO1, KRT8) tissue- or cell-specific expression in the Protein Atlas database (Supplementary Table 1). Among all 74 DB TF target genes, 30 have reproductive specific expression, 28 have brain specific expression, 18 have blood cell specific expression, and 16 have mitochondria specific expression or localization.

Table 2. DB TF target genes with differential binding affinity by DB TFs in modern human vs. Neanderthal proximal and semi-distal promoter regions.

Gene	Gene description	Single-cell RNA cell specific	Count
FARS2	Phenylalanyl-tRNA synthetase 2, mitochondrial	astrocytes, early spermatids, excitatory neurons, inhibitory neurons, late spermatids, microglial cells, oligodendrocyte precursor cells, oligodendrocytes	143
TEX19	Testis expressed 19	spermatocytes, spermatogonia	60
NARF	Nuclear prelamin A recognition factor	erythroid cells, late spermatids	56
SLC6A11	Solute carrier family 6 member 11	astrocytes, excitatory neurons, inhibitory neurons, muller glia cells, oligodendrocyte precursor cells	54
LSM5	LSM5 homolog, U6 small nuclear RNA and mRNA degradation associated		49
FAM172A	Family with sequence similarity 172 member A	astrocytes, excitatory neurons, inhibitory neurons, microglial cells, oligodendrocyte precursor cells, oligodendrocytes	47

WNT7A	Wnt family member 7A	cytotrophoblasts, extravillous trophoblasts, pancreatic endocrine cells, syncytiotrophoblasts	43
EXO1	Exonuclease 1	early spermatids, erythroid cells, extravillous trophoblasts, plasma cells, spermatocytes, spermatogonia	42
KRT8	Keratin 8	distal enterocytes, ductal cells, exocrine glandular cells, extravillous trophoblasts, pancreatic endocrine cells, paneth cells, proximal enterocytes, syncytiotrophoblasts	35
ZNF37A	Zinc finger protein 37A		32
TIGD6	Tigger transposable element derived 6	spermatocytes	27
CSTF3	Cleavage stimulation factor subunit 3		21
GDPGP1	GDP-D-glucose phosphorylase 1		18
HLA-DQB1	Major histocompatibility complex, class II, DQ beta 1	b-cells, langerhans cells, macrophages, monocytes	16
LSM10	LSM10, U7 small nuclear RNA associated		15
ZNF132	Zinc finger protein 132	cardiomyocytes, cone photoreceptor cells, early spermatids	14
CCDC197	Coiled-coil domain containing 197		13
AK4	Adenylate kinase 4	hepatocytes, proximal tubular cells	13
DDX43	DEAD-box helicase 43	spermatocytes, spermatogonia	11
AKAP13	A-kinase anchoring protein 13		9
ZC3H3	Zinc finger CCCH-type containing 3		9
ZCCHC9	Zinc finger CCHC-type containing 9		9
MFAP3L	Microfibril associated protein 3 like	late spermatids	8
IFITM2	Interferon induced transmembrane protein 2	adipocytes, nk-cells	8
GPA33	Glycoprotein A33	distal enterocytes, intestinal goblet cells, paneth cells, proximal enterocytes, undifferentiated cells	8
MMEL1	Membrane metalloendopeptidase like 1	early spermatids, late spermatids, paneth cells	8
PIGP	Phosphatidylinositol glycan anchor biosynthesis	spermatocytes	7

	class P		
COA6	Cytochrome c oxidase assembly factor 6		7
PCNX4	Pecanex 4	dendritic cells, endometrial ciliated cells	7
PHGDH	Phosphoglycerate dehydrogenase	distal tubular cells, exocrine glandular cells	6
NPTXR	Neuronal pentraxin receptor	cardiomyocytes, excitatory neurons, inhibitory neurons, oligodendrocyte precursor cells	6
IMMT	Inner membrane mitochondrial protein	cardiomyocytes	6
CDC37L1	Cell division cycle 37 like 1		5
SMC2	Structural maintenance of chromosomes 2	endometrial ciliated cells	5
AGR2	Anterior gradient 2, protein disulphide isomerase family member	club cells, distal enterocytes, gastric mucus-secreting cells, intestinal goblet cells, paneth cells, undifferentiated cells, urothelial cells	4
SLC8A3	Solute carrier family 8 member A3	bipolar cells, excitatory neurons, inhibitory neurons, oligodendrocyte precursor cells	4
AKT1S1	AKT1 substrate 1	syncytiotrophoblasts	4
HNRNPUL1	Heterogeneous nuclear ribonucleoprotein U like 1		4
RNASEH2A	Ribonuclease H2 subunit A	cytotrophoblasts, extravillous trophoblasts	3
PRR18	Proline rich 18	oligodendrocytes	2
KCNQ3	Potassium voltage-gated channel subfamily Q member 3	excitatory neurons, inhibitory neurons, microglial cells	2
ITGA1	Integrin subunit alpha 1	adipocytes, endothelial cells, hepatocytes, sertoli cells, smooth muscle cells	2
GRIN2A	Glutamate ionotropic receptor NMDA type subunit 2A	excitatory neurons, inhibitory neurons	2
TACC3	Transforming acidic coiled-coil containing protein 3	early spermatids, hofbauer cells, late spermatids, spermatocytes	2
EIF2A	Eukaryotic translation initiation factor 2A		2

MPHOSPH9	M-phase phosphoprotein 9	cone photoreceptor cells, excitatory neurons, inhibitory neurons	2
F7	Coagulation factor VII	hepatocytes	2
ZNF211	Zinc finger protein 211		2
KLHL24	Kelch like family member 24	oligodendrocytes	2
DKKL1	Dickkopf like acrosomal protein 1	early spermatids, late spermatids	2
STPG2	Sperm tail PG-rich repeat containing 2	astrocytes, excitatory neurons, inhibitory neurons, microglial cells, oligodendrocyte precursor cells, oligodendrocytes	1
PLD3	Phospholipase D family member 3	hofbauer cells	1
MZT1	Mitotic spindle organizing protein 1	erythroid cells	1
EVI5	Ecotropic viral integration site 5	oligodendrocytes	1
DGCR2	DiGeorge syndrome critical region gene 2		1
GABRA5	Gamma-aminobutyric acid type A receptor subunit alpha5	excitatory neurons, inhibitory neurons	1
CYP4F12	Cytochrome P450 family 4 subfamily F member 12	distal enterocytes, gastric mucus-secreting cells, paneth cells, proximal enterocytes	1
RNF215	Ring finger protein 215	bipolar cells, cone photoreceptor cells, rod photoreceptor cells	1
ADARB2	Adenosine deaminase RNA specific B2 (inactive)	inhibitory neurons, oligodendrocyte precursor cells	1
MYL12A	Myosin light chain 12A	cardiomyocytes	1
RUVBL2	RuvB like AAA ATPase 2	early spermatids, late spermatids, respiratory ciliated cells, spermatocytes	1
TRIM2	Tripartite motif containing 2	oligodendrocytes	1
CADPS	Calcium dependent secretion activator	bipolar cells, cone photoreceptor cells, excitatory neurons, inhibitory neurons, oligodendrocyte precursor cells	1
PCK1	Phosphoenolpyruvate carboxykinase 1	distal enterocytes, distal tubular cells, hepatocytes, proximal enterocytes, proximal tubular cells	1
NGDN	Neuroguidin		1

ABCA8	ATP binding cassette subfamily A member 8	fibroblasts, granulosa cells, leydig cells, microglial cells, oligodendrocytes, skeletal myocytes	1
POGZ	Pogo transposable element derived with ZNF domain	oligodendrocytes	1
PDIA6	Protein disulfide isomerase family A member 6	extravillous trophoblasts, plasma cells	1
ACSL1	Acyl-CoA synthetase long chain family member 1	early spermatids, hepatocytes, late spermatids	1
SUPV3L1	Suv3 like RNA helicase		1
DCXR	Dicarbonyl and L-xylulose reductase	distal tubular cells, hepatocytes, proximal tubular cells	1
DPYSL5	Dihydropyrimidinase like 5	cone photoreceptor cells, late spermatids, oligodendrocytes	1
OMA1	OMA1 zinc metallopeptidase	oligodendrocytes	1
NME1	NME/NM23 nucleoside diphosphate kinase 1		1

DB TFs are highly expressed in immune cells

Data derived from the bulk RNA-Seq experiments of the FANTOM5 project was used to generate a cluster map of DB TF expression within the 100 tissues with the highest aggregate expression, presented as Figure 1A. Relative aggregate expression for individual tissues is presented as bars in Figure 1B and shows that within the 100 tissues those with the highest expression are immune cell types/tissues. The immune cell types with highest aggregate DB TF expression are mast cells, eosinophils, CD4+ T cells, CD8+ T cells, Basophils, $CD14^{+}/CD16^{-}$ monocytes, and natural killer cells (Supplementary Figure 1); and of the 20 cell types/tissues with the highest aggregate expression, 16 are immune.

DB TFs coexpress in neural and immune tissues

Expression data for DB TFs in the FANTOM 5 dataset was extracted for cluster analysis (Figure 1A). Tissues of several types cluster together by their expression of DB TFs, with the largest clusters being neural (30 brain and eye/retinal tissues) and immune (28 immune cells/tissues) (Figure 1B). Likewise, based on their tissue expression in the FANTOM data set, the DB TFs have been clustered. Data on tissue group specificity (brain and retina; male/female) from the Human Protein Atlas (bulk and single-cell RNA-Seq), status as a development related gene (homeobox, forkhead box, and SRY-related HMG-box), and above 90th quantile expression in immune cell types as defined by the Database of Immune Cell eQTLs (DICE), was analyzed and included as Figure 1C. This data shows two distinct top-level clusters: the larger cluster contains 89 genes, the majority of which are known to have a role in development (78/89) and specific expression in a brain or retina tissue (72/89); while the smaller cluster of 17 genes all show specific expression in blood and immune cells.

Within the larger cluster of 89 developmental and brain specific genes there is a subcluster of 12 genes which whose expression is most limited to FANTOM 5 neural tissues: NKX6-2, CUX2, NR2E1, POU3F2, POU6F2, EMX1, VAX1, POU3F3, DLX1, DLX2, OTX1, POU3F1. All of these genes are identified in the Protein Atlas as being enriched or enhanced in the brain or retina. GO analysis of the 12 genes with the PANTHER-based geneontology.org web server [51,52] shows the top 10 enriched biological process terms: ‘cerebral cortex GABAergic interneuron fate commitment’ (GO:0021893; 100x enrichment), 'positive regulation of amacrine cell differentiation' (GO:1902871; 100x), 'forebrain ventricular zone progenitor cell division' (GO:0021869; 100x), 'negative regulation of photoreceptor cell differentiation' (GO:0046533; 100x), 'regulation of photoreceptor cell differentiation' (GO:0046532; 100x), 'regulation of amacrine cell differentiation' (GO:1902869; 100x), 'neuroblast differentiation' (GO:0014016; 100x), 'negative regulation of oligodendrocyte differentiation' (GO:0048715; 100x), 'cerebral cortex GABAergic interneuron

differentiation' (GO:0021892; 100x), and 'negative regulation of glial cell differentiation' (GO:0045686; 100x).

Additionally, within the brain tissues cluster we observe sub-clusters of fetal and newborn tissues (Figure 1D). Overall, DB TFs show the highest expression in medial frontal gyrus (newborn), medial temporal gyrus (adult), parietal lobe (fetal), occipital lobe (fetal), eye (fetal), pineal gland (adult), temporal lobe (fetal), retina (adult), occipital cortex (newborn), parietal lobe (newborn), dura mater (adult), medial temporal gyrus (newborn), and spinal cord (fetal) (Supplementary Figure 1). In a similar fashion, the cluster with high expression in blood and immune cells/tissues is comprised of 17 genes: FOS, RORA, FOXK1, FOXK2, IRF3, NFYA, TBP, MAFG, POU2F1, ARID5A, RARA, JUN, CUX1, FOXP1, MEF2C, MAFB, and RXRA All of these genes have expression ⩾90th quantile of all genes in at least one immune cell type as cataloged in the Database of Immune Cell eQTLs (DICE) [53]. GO analysis of the 17 genes (geneontology.org web server) shows enrichment for immune related biological process terms: ‘CD4-positive, alpha-beta T cell differentiation involved in immune response’ (GO:0002294; 76x enrichment), ‘T-helper cell differentiation’ (GO:0042093; 76x), and ‘alpha-beta T cell differentiation involved in immune response’ (GO:0002293; 73x), among others.

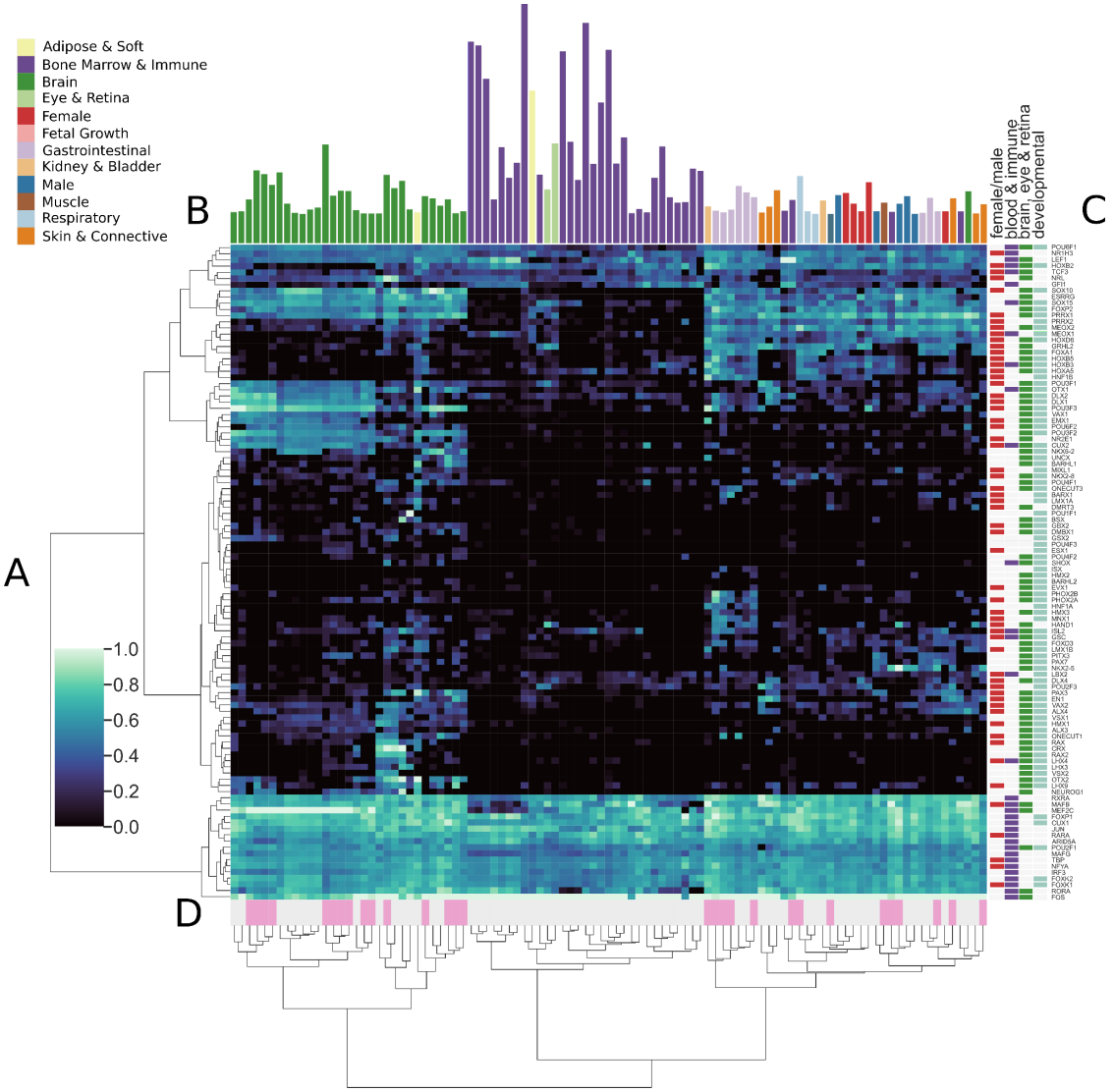

Figure 1. Cluster map of expression of transcription factors displaying variable binding in modern human vs. Neanderthal. (**A**) Expression data derived from the FANTOM 5 dataset is used to cluster 106 DB transcription factors in 100 tissues. Expression data was not available for DUXA and PROP1 genes. (**B**) Aggregate expression for each tissue across all genes is presented as bars, and bars/tissues are colored by tissue group. Distinct clusters of brain and eye/retina tissues as well as blood and immune cells/tissues are observed. Barplot with tissue names is presented in Supplementary Figure 1. (**C**) TF genes are labeled with colored bars corresponding to their identification as homeobox, SOX, or FOX genes (developmental), above 90th quantile expression in an immune cell type as defined by DICE database (blood & immune) (Schmiedel et al. 2018), or to have enriched or enhanced expression by tissue as defined by proteinatlas.org (female/male; brain, eye & retina)(Uhlen et al. 2015). (**D**) Life stage for each tissue is described, where white and pink represents adult and fetal/newborn tissues, respectively. Expression values (TPM) have been normalized by log transformation and each tissue column has been adjusted on a zero to one scale.

DB TFs have developmental time-point related expression profiles in brain

Cluster analysis of Allen Brain Atlas bulk RNA-Seq expression data for DB TF genes revealed distinct clusters of time-point specific expression (Figure 2B). Specifically, we observe clusters of expression

for brain tissues at different grouped time points: pcw (8 to 37 weeks post conception), early (4 months to 4 years), and late (8 years to 40 years). Nearly all of the pre-birth (pcw) brain tissues cluster together, without admixture with the other grouped timepoints. The exception is a small cluster where 'cerebellar cortex' from all three groups is joined with 'pcw cerebellum', 'pcw upper (rostral) rhombic lip', and ‘mediodorsal nucleus of thalamus’ of all three groups. All 15 pre-birth cortical brain tissues form a cluster. Similarly, a large cluster contains all of the other early and late tissues which primarily segregate into two age-related subclusters. Those genes comprising the cluster with highest aggregate expression in the data set (module 1 genes: NFYA, TCF3, MEF2C, SOX10, CUX2, TBP, FOXP1, IRF3, SOX15, RARA, FOXK2, NKX6-2, FOXK1, FOS, JUN, POU6F1, POU3F2, MAFG, POU3F3, MAFB, and CUX1) show a dynamic and cyclic pattern of expression during the pre-birth timepoints which then stabilizes in infancy and trends downwards in early childhood (Figure 2A).

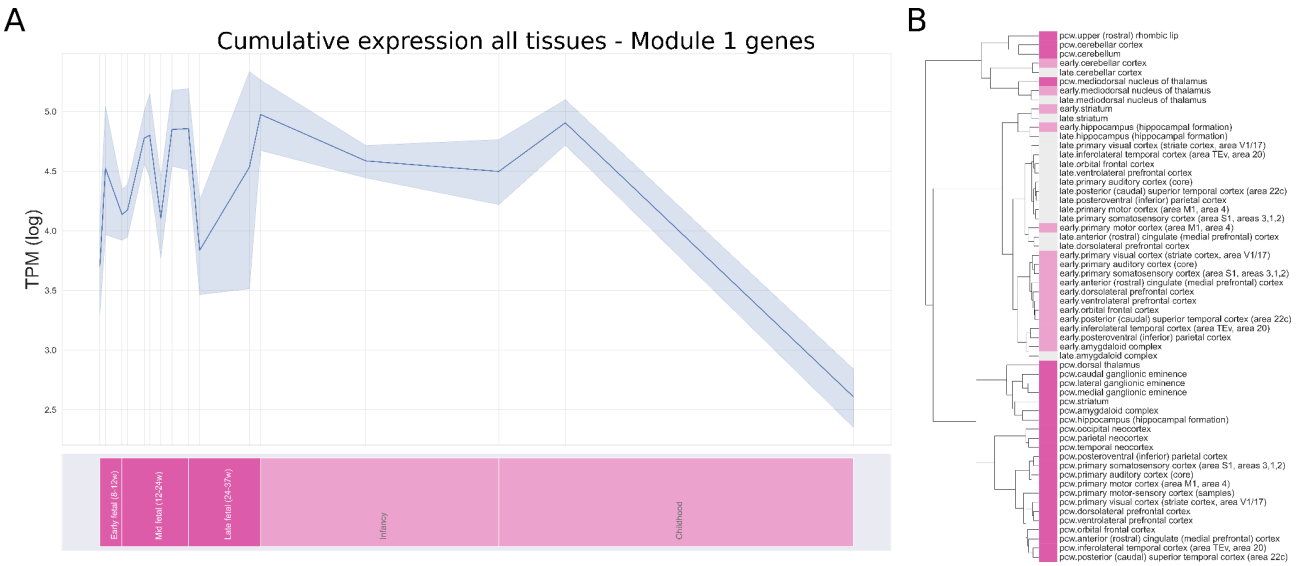

Figure 2. Expression of modern human vs Neanderthal DB TFs in developing brain tissue. Expression data from Allen Brain Atlas [54] was extracted as RPKM values for 26 unique tissues across 31 timepoints, converted to TPM and log transformed, and used to calculate a cluster map. A) Aggregate expression of the cluster of genes with highest total expression across all tissues (module 1), from 8 to 144 weeks post conception. B) Clustering of brain tissues. Colors represent development stages, for both figures: 8 to 37 weeks post conception (pcw, magenta), 4 months to 4 years (early, pink), and 8 years to 40 years (late, white).

Single-nucleus RNA-Seq of cortical cells reveals differential expression of DB TFs

Analysis of Allen Brain Atlas single-nucleus RNA-Seq data from 49,417 cortical brain cells [55] was performed to identify which of the MH-Neanderthal DB TF genes may play a functional role in specific annotated brain regions, layers, and cell types. Analyses in the original paper have defined metadata indicating cell class (excitatory, inhibitory, and non-neuronal), layer (L1–L6), and predicted type (e.g., astrocyte, microglia, endothelial, etc.), as well as brain regions (Figure 3A). Cell groups defined by the intersection of class, layer, and region were analyzed for differentially expressed genes (DEGs). Several DB TFs were DEGs in specific categories, occurring in nearly all cases in glutamatergic neurons, usually in layers L4–L6, and across nearly all brain regions. Categories where four or more DB TFs were DEGs are presented in Table 3, and included cut-like homeobox 1 (CUX1), cut-like homeobox 2 (CUX2), estrogen-related receptor gamma (ESRRG), forkhead box protein 1 (FOXP1), forkhead box protein 2 (FOXP2), Myocyte Enhancer Factor 2C (MEF2C), POU Class 6 Homeobox 2 (POU6F2), paired related homeobox 1 (PRRX1), and

RAR-related orphan receptor alpha (RORA). In all categories POU6F2 was a DEG, and in all but two either FOXP2 or FOXP1 (or both) were DEGs. FOXP2 was present as a DEG in excitatory L4 cells in A1C (primary visual cortex), MTG (medial temporal gyrus), S1lm (lower limb somatosensory cortex), and S1ul (upper limb somatosensory cortex).

Table 3. DB TFs which co-occur as DEGs in specific regions and cell types of Allen Brain Atlas scRNA-Seq of cerebral cortex

Layer	Class	Region	DEG DB TF genes
L4	Glutamatergic	A1C	FOXP2, FOXP1, POU6F2, MEF2C
L4	Glutamatergic	MTG	FOXP2, PRRX1, POU6F2, ESRRG
L4	Glutamatergic	S1lm	FOXP2, PRRX1, POU6F2, ESRRG
L4	Glutamatergic	S1ul	FOXP2, FOXP1, RORA, POU6F2, MEF2C
L4	Glutamatergic		FOXP1, CUX2, RORA, POU6F2, CUX1, MEF2C
L4		IT	FOXP1, CUX2, RORA, POU6F2, CUX1, MEF2C
L4ab	Glutamatergic	V1C	CUX2, RORA, POU6F2, CUX1, MEF2C
L4ab			CUX2, RORA, POU6F2, CUX1
L5	Glutamatergic	S1lm	FOXP2, PRRX1, POU6F2, ESRRG
L5	Glutamatergic	V1C	FOXP2, FOXP1, RORA, POU6F2, MEF2C
L6a			FOXP2, PRRX1, POU6F2, ESRRG

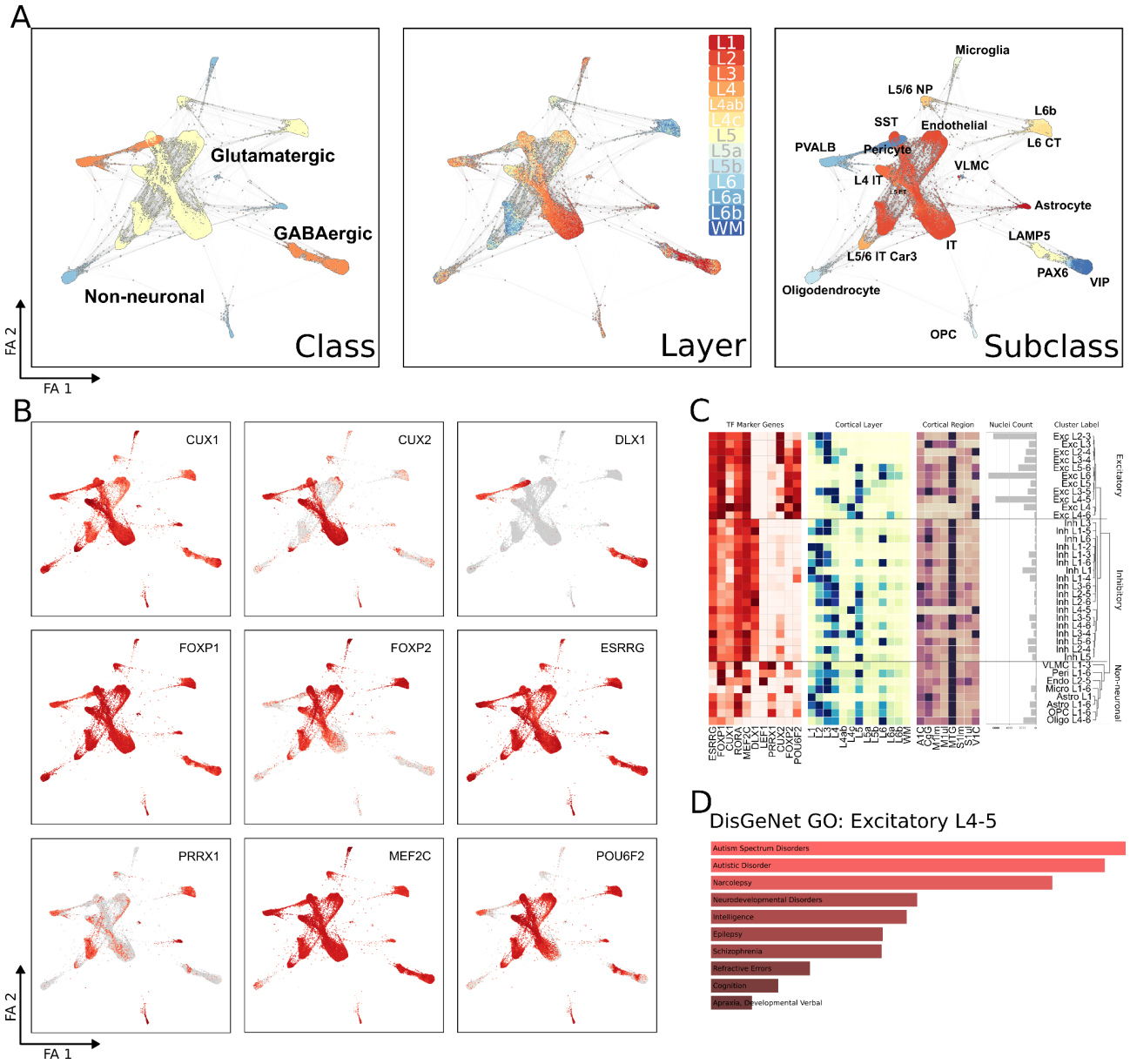

Figure 3. Analysis of single-nucleus RNA-Seq data from brain cortical regions. (**A**) Force-directed graph (FA) visualization of Allen Brain Atlas single-nucleus RNA-Seq data from 49,417 brain cortex cells; brain excitatory/inhibitory status, cortical layer, and predicted cell type (subclass) and are mapped from annotations in the Allen Brain Atlas sample data [55]. (B) Expression of DB TFs which are DEGs mapped onto FA visualization. C) On the x-axis: DB TF DEGs, cortical layer, and cortical region; on the y-axis: cell class/type and cortical layer, with the dendrogram indicating expression relationships. Color intensity indicates normalized expression (0–1 scale). Nuclei count indicates the number of nuclei for that cell class/type at the indicated layer. D) DisGeNET disease ontology terms enriched for the top 100 marker genes for L4–5 excitatory cells. ET, extratelencephalic/pyramidal tract; IT, intratelencephalic; NP, near-projecting; OPC, oligodendrocyte precursors; VLMC, vascular leptomeningeal cell; SST, LAMP5, PAX6, VIP, PVALB, labels based on expression of corresponding gene.

Gene ontology analysis of L4 excitatory neurons

In the Allen Brain Atlas cortical cell snRNA-Seq data, the multiple DB TFs were consistently DEGs in both layer 4 and excitatory cell groups, and their intersection. Gene ontology analysis of the top 100 marker genes of L4 excitatory cells using the DisGeNet ontology within the Enrichr tool [56] showed the top over enriched terms ‘autism spectrum disorders’, ‘autistic disorder’, ‘narcolepsy’, ‘neurodevelopmental disorders’, ‘intelligence’, ‘epilepsy’, ‘schizophrenia’, ‘refractive errors’, ‘cognition’, and ‘apraxia, developmental verbal’ (Figure 3D). The top 10 biological process ontology over enriched terms for L4 excitatory cells were: ‘corticospinal neuron axon guidance’ (GO:0021966; >100x enrichment), ‘corticospinal tract morphogenesis’ (GO:0021957; 91x), ‘ventricular cardiac muscle cell differentiation’ (GO:0055012; 42x), ‘synaptic membrane adhesion’ (GO:0099560; 35x), 'central nervous system projection neuron axonogenesis' (GO:0021952: 30x), 'synaptic transmission, GABAergic' (GO:0051932: 29x), 'positive regulation of excitatory postsynaptic potential' (GO:2000463: 28x), 'negative regulation of smooth muscle cell migration' (GO:0014912: 27x), 'positive regulation of dendrite morphogenesis' (GO:0050775: 25x), and 'modulation of excitatory postsynaptic potential' (GO:0098815: 25x).

In GO analysis of cellular component, the top 10 over enriched terms for L4 excitatory cells were: ‘anchored component of presynaptic membrane' (GO:0099026; 61x enrichment), ‘presynaptic cytosol' (GO:0099523; 47x), ‘NMDA selective glutamate receptor complex' (GO:0017146; 47x), ‘intrinsic component of presynaptic active zone membrane' (GO:0098945; 45x), ‘node of Ranvier' (GO:0033268; 42x), ‘integral component of presynaptic active zone membrane' (GO:0099059; 40x), ‘GABA-A receptor complex' (GO:1902711; 34x), ‘GABA receptor complex' (GO:1902710; 30x), ‘presynaptic active zone membrane' (GO:0048787; 27x), and ‘intrinsic component of presynaptic membrane' (GO:0098889; 22x).

In GO analysis of molecular function, the top 10 over enriched terms for L4 excitatory cells were: 'transmembrane receptor protein tyrosine phosphatase activity' (GO:0005001; 50x enrichment), 'transmembrane receptor protein phosphatase activity' (GO:0019198; 50x), 'GABA-gated chloride ion channel activity' (GO:0022851; 49x), 'inhibitory extracellular ligand-gated ion channel activity' (GO:0005237; 42x), 'ligand-gated anion channel activity' (GO:0099095; 35x), 'GABA-A receptor activity' (GO:0004890; 34x), 'syntaxin-1 binding' (GO:0017075; 32x), 'GABA receptor activity' (GO:0016917; 29x), 'transmitter-gated channel activity' (GO:0022835; 17x), and 'transmitter-gated ion channel activity' (GO:0022824; 17x).

DB TFs and Human Accelerated Regions

Doan et al. have identified and cataloged a list of genes in close proximity to human accelerated regions (HARs; elevated divergence in humans vs. other primates) as well as those which have direct interaction with a HAR through chromatin conformation changes, e.g., enhancers [57]. In the set of 110 DB TF genes, 28 were found to be HAR-associated. Fisher's exact test analysis revealed that over-representation of HAR-association in DB TFs versus all JASPAR TFs was not significant (OR 0.98; p-value 0.57). However, compared to all protein-coding genes, the JASPAR TFs as a group are enriched for HAR-association (OR 3.27; p-value 6.73×10^{-26}).

TFBSFootprinter availability

The TFBSFootprinter tool is available for installation via Conda making it available for most computing OS systems (https://anaconda.org/thirtysix/tfbs-footprinting3). The TFBSFootprinter tool is also available for use as a Python library (https://pypi.org/project/TFBS-footprinting3/) and subsequently can be easily installed to a Linux system using a single command 'pip install TFBS-footprinting3'. Due to size considerations, supporting experimental data for both human and non-human species is downloaded on demand on first usage. Documentation on background,

usage, and options is available both within the program and more extensively online (tfbs-footprinting.readthedocs.io).

Experimental Datasets

Experimental data from a total of six databases were incorporated into the TFBSFootprinter algorithm (Figure 4A). Data from the relevant datasets were pre-processed to generate score distributions with which putative TFBS predictions could later be compared, as described in the Methods. Each dataset allows for scoring of transcription-relevant markers in or near putative regulatory elements identified by PWM analysis: co-localization with ChIP-Seq metacluster, cap analysis of gene expression (CAGE) peak, ATAC-Seq peak, or CpG island; correlation of expression between predicted TF and gene of interest; co-localization of eQTL and effect on expression of target gene; measure of conservation in related vertebrate species (Figure 4C). The simplicity of this piece-wise approach allows for easy inclusion of additional TFBS relevant data in the future.

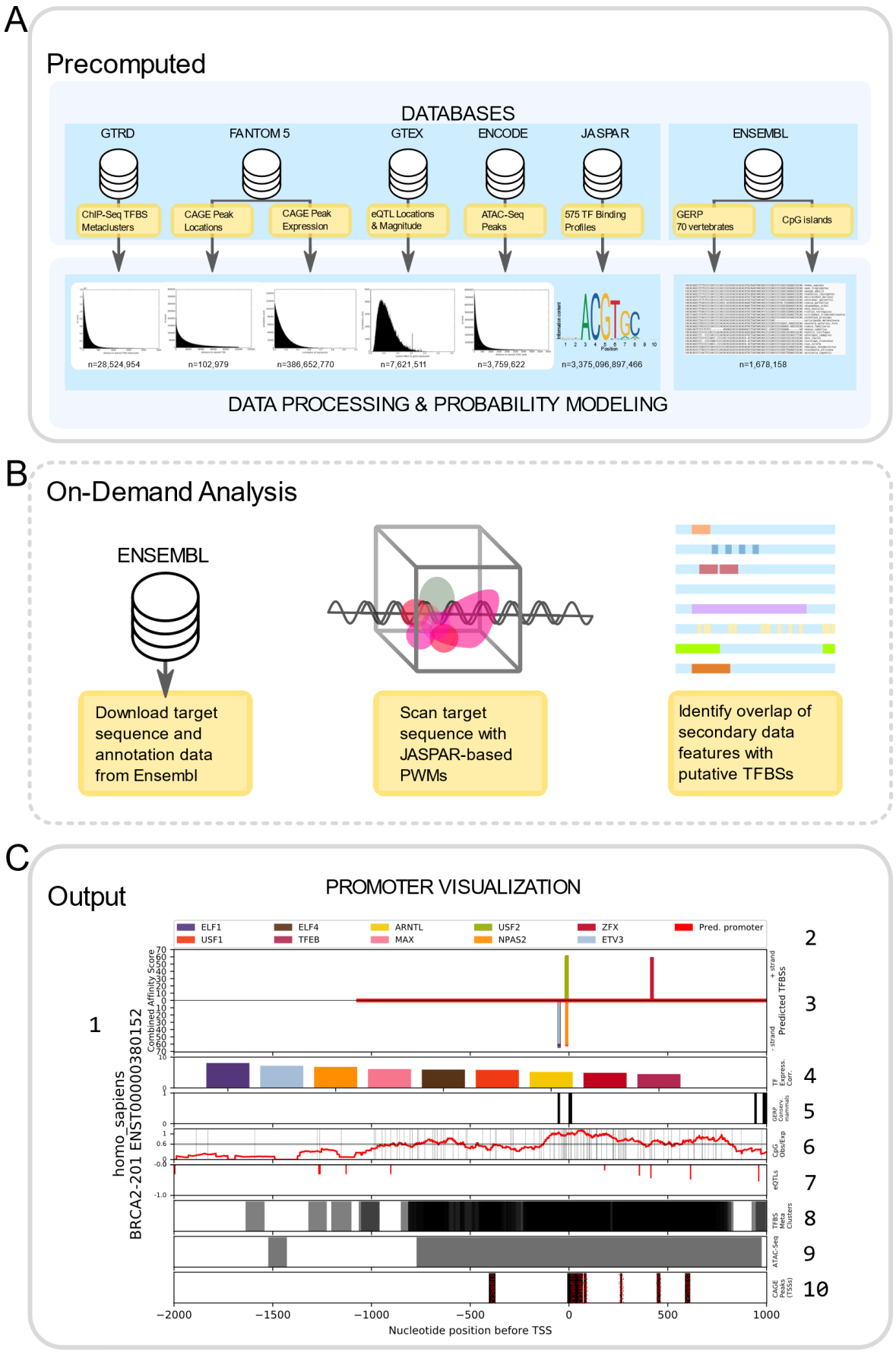

Figure 4. Outline of the datasets used in TFBSFootprinter. (**A**) A total of six empirical datasets are used to support computational prediction of TFBSs in the TFBSFootprinter tool. Experimental data is pre-processed to generate score

distributions from which probability scores can be applied to putative TFBSs (n values indicate number of elements used to compute distributions). PWMs are created from JASPAR PFMs, and a segment of DNA is extracted from the Ensembl database for a region surrounding a target TSS of an Ensembl annotated transcript ID. Log-likelihood scoring is performed for six parameters on the target sequence. (**B**) A user-defined target promoter sequence and annotation data are downloaded, and JASPAR PWMs are created on-demand. PWM analysis of the promoter sequence generates putative TFBSs hits which are then compared with elements from pre-processed experimental datasets and scored using pre-generated probability scores relevant for the target genome. (**C**) The outputs of the TFBSFootprinter analysis are a publication-ready scalable vector graphics file (.svg), a table of results including predicted TFBSs names, locations, and scoring for each metric (not pictured), as well as individual files containing sequences and annotations. (C1) HUGO Gene Nomenclature Committee (HGNC)-based identifier + Ensembl transcript ID. (C2) Color coded legend of the top 10 transcription factors predicted to bind to this promoter. (C3) Graphical representation of promoter of transcript, predicted binding sites are indicated by bars. Bar height represents the combined affinity score (by default, a summation of log-likelihood scores from each transcription-relevant dataset, or alternatively a subset of datasets defined by user). Positive y-axis indicates binding on the positive (sense) strand and negative y-axis represents negative (anti-sense) strand. (C4) Score of the correlation of expression between each top predicted TF and the target gene. (C5) Highly conserved regions of 70-mammal alignment analyzed as determined by GERP analysis (black bars). (C6) Vertical lines represent CpG locations. The red line describes the CpG ratio of the promoter sequence over a 100 nt window. (C7) Genetic variants identified in the GTEx database to have an effect on the target gene's expression (eQTLs). Green indicates positive impact on expression (positive y-axis) and red indicates negative (negative y-axis). (C8) TFBS metaclusters identified in the GTRD database (grey bars). (C9) Cell-type agnostic ATAC-Seq peaks (open chromatin) retrieved from the ENCODE database (grey bars). (C10) CAGE peaks indicating TSSs identified in the FANTOM database (black bars). Nucleotide positions at the bottom are relative to the Ensembl-defined transcription start site of the target transcript, and apply to C3 and C5-C10.

Inclusion of empirical datasets improves TFBSFootprinter accuracy

The performance of both individual datasets and combinations of datasets, in the identification of experimentally verified functional TFBSs and TFBS ChIP-Seq peaks, was tested by receiver operating characteristic (ROC) analysis (Figure 5, Supplementary Figure 2, Supplementary Figure 3, Supplementary Figure 4, Figure 6, Supplementary Figure 5). Four different benchmarking approaches were used. Depending on the benchmarking approach used, different combinations of transcription-relevant features produced the best ROC scores. We observed that when using all available features TFBSFootprinter consistently outperformed the PWM, and that for the greater majority of TFs tested the best TFBSFootprinter model outperformed the best DeepBind model (benchmark 1, 60/65; benchmark 2, 28/40; benchmark 3, 13/14; and benchmark 4, 11/14). In the majority of cases a subset of the available transcription-relevant data produced the optimal model (Supplementary Figure 2). In addition, the features which were most frequently observed as

components of the best models (Figure 5A, Figure 6A, Supplementary Figure 3A, Supplementary Figure 4A) varied by benchmark.

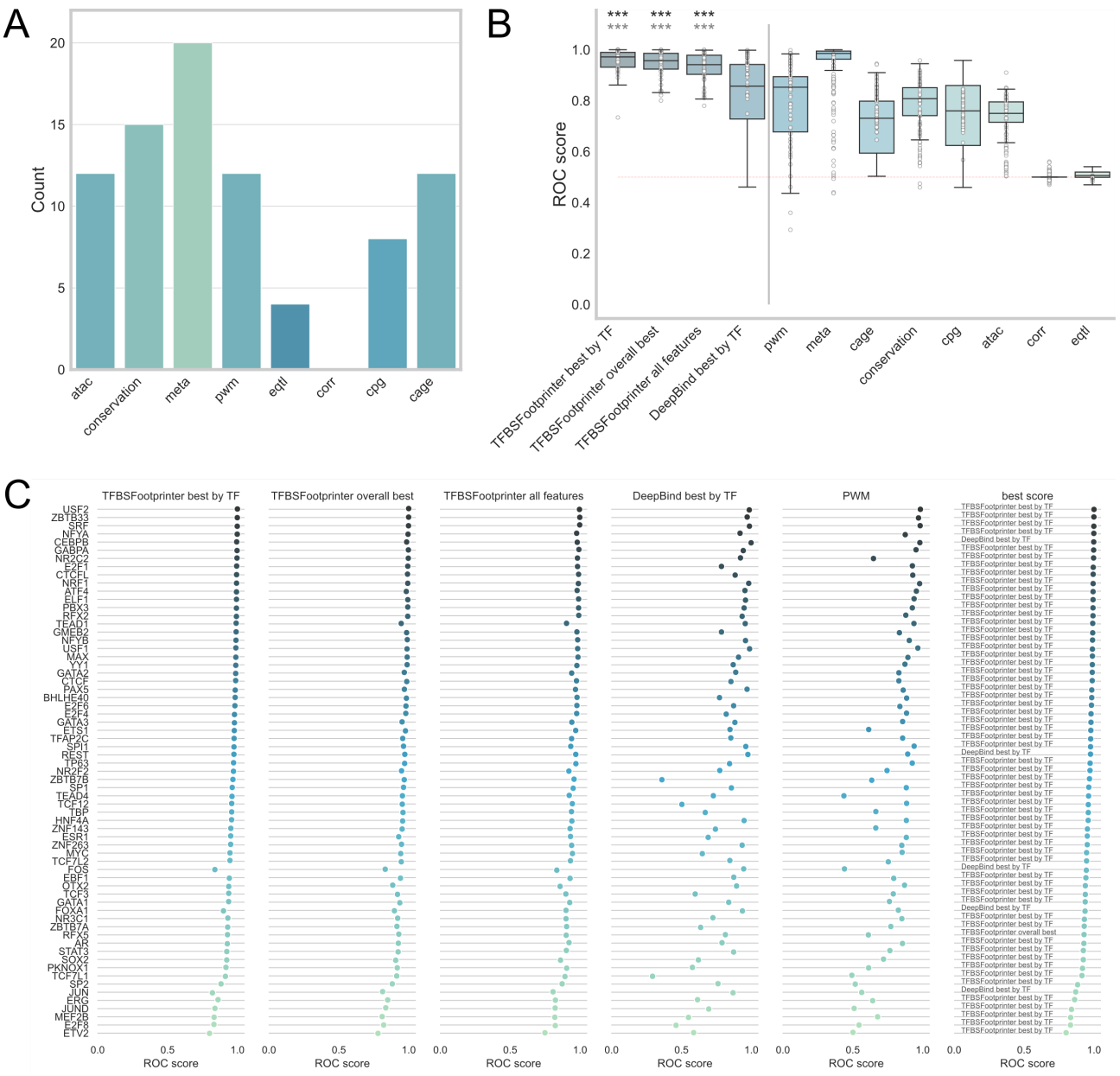

Figure 5. ROC Analysis model performance in the identification of TFBS ChIP-Seq peaks — strong and distal binding. ROC analysis was performed for TFs modeled by both JASPAR and DeepBind using associated GTRD ChIP-Seq peaks. Peaks with fold-enrichment ≥50x located within -900 and +200 bp relative to an Ensembl-defined protein-coding transcript as true positives, and paired sites 1,000 bp upstream as true negatives. ROC analysis was only performed on TFs which had at least 100 true positives and 100 true negatives. Each true positive/negative segment analyzed was 50 nucleotides long, with the highest TFBS score kept as representative. (**A**) Barplot of the frequency of experimental data type in the top 20 performing TFBSFootprinter models. (**B**) Boxplot of ROC scores for TFBSFootprinter and DeepBind for 64 TFs. ROC scores were also calculated based on using individual experimental metrics to show how well each contributes to accuracy of the combined model. (**C**) ROC scores for each individual TF tested, for each primary TFBS prediction model under study. The best scoring model among all is named for each TF (right). TFBSFootprinter best by

TF (specific models and ROC score shown in Supplementary Figure 2), based on using the highest ROC score achieved by some combination of experimental data models; TFBSFootprinter overall best, based on using the combination of experimental data models which had the best average ROC score across all TFs analyzed; DeepBind best by TF, based on using the higher ROC score of the SELEX or ChIP-Seq DeepBind models. Black asterisks denote significant difference of first four models with 'pwm' model and gray asterisks denote significant difference between TFBSFootprinter-based models and DeepBind model, as determined by related t-test; * p-value<0.05, ** p-value <0.005, and *** p-value <0.0005.

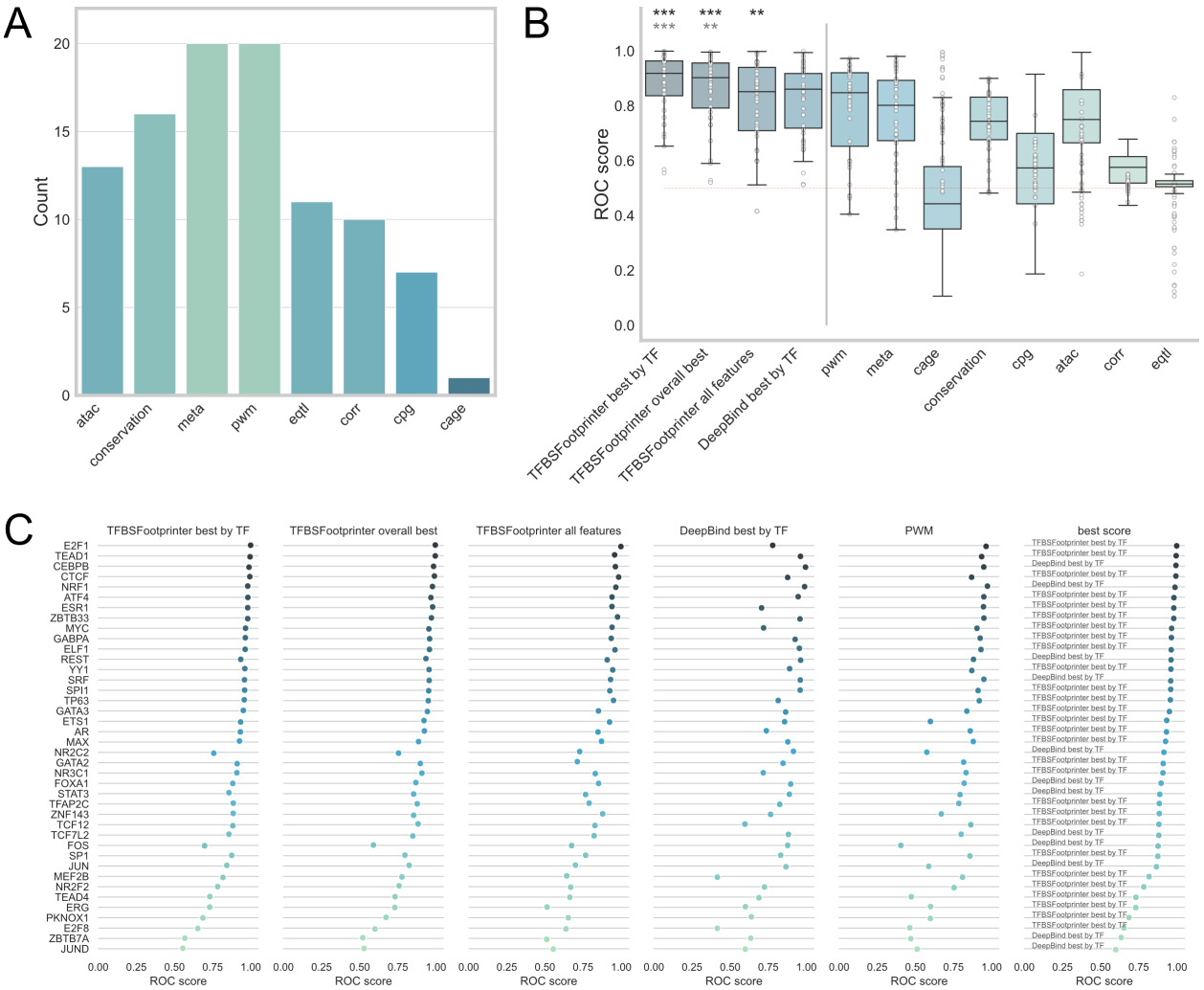

Figure 6. ROC Analysis model performance in the identification of TFBS ChIP-Seq peaks — strong and weak binding. ROC analysis was performed for TFs modeled by both JASPAR and DeepBind using associated GTRD ChIP-Seq peaks located within -900 and +200 bp relative to an Ensembl-defined protein-coding transcript, with fold-enrichment ≥50x as true positives, and those with fold-enrichment ≤2x as true negatives. ROC analysis was only performed on TFs which had at least 100 true positives and 100 true negatives. Each true positive/negative segment analyzed was 50 nucleotides long, with the highest TFBS score kept as representative. (**A**) Barplot of the frequency of experimental data type in the top 20 performing TFBSFootprinter models. (**B**) Boxplot of ROC scores for TFBSFootprinter and DeepBind for 40 TFs. ROC scores were also calculated based on using individual experimental metrics to show how well each contributes to accuracy of the combined model. (**C**) ROC scores for each individual TF tested, for each primary TFBS prediction model under study. The best scoring model among all is named for each TF (right). TFBSFootprinter best by TF (specific models and ROC score shown in Supplementary Figure 5), based on using the highest ROC score achieved by some combination of experimental data models; TFBSFootprinter overall best, based on using the combination of experimental data models which had the best average ROC score across all TFs analyzed; DeepBind best by TF, based on using the higher ROC score of the SELEX or ChIP-Seq DeepBind models. Black asterisks denote significant difference of first four models with 'pwm' model and gray asterisks denote significant difference between TFBSFootprinter-based

models and DeepBind model, as determined by related t-test; * p-value<0.05, ** p-value <0.005, and *** p-value <0.0005.

Discussion

In this study we have introduced and tested a new method for prediction of transcription factor binding sites. The approach levies nine separate transcription-relevant data to augment prediction of functional TFBSs beyond the classical PWM. As a proof of use, we have used the tool to analyze differences in regulatory regions observed in comparison of modern human and Neanderthal DNA. The results align with and complement previous studies on Neanderthal DNA, and with studies on brain development and evolution. While the current study focuses on human DNA, it is important to note that the TFBSFootprinter tool is built to also make predictions in many other species, based on its integration of Ensembl data.

Regulatory changes drive evolution

Current knowledge points to the fact that, across the tree of life, prior to and during major radiations of new species there comes a commensurate increase in the number of regulatory genes. Comparative genomic analyses suggest that at the time of eukaryogenesis, or the origin of the last eukaryotic ancestor (LECA), a significant increase in novel TF classes occurred [58]. Eukaryogenesis is one of the major transitions of life on Earth, with an explosion of diversity owing to the endosymbiotic synthesis of an energy producing α-proteobacteria mitochondrion-progenitor within an archaeon [59]. If the new inhouse energy source could be described as the engine driving the diversity arising from LECA, the complementary argument could be made that it was TF proteins which did the steering. Later, prior to the colonization of land by plants, there was an increase in the TF families of the ancestral aquatic streptophytes which were then already present in the first land plants [60]. Again, prior to the radiation of

bilaterian (multicellular) metazoans an increase in transcription factor families occurred [61], roughly quadrupling in ratio [62]. Arguably, these increases in TF family numbers immediately prior to major transitions in life could be pointed to as evidence of the critical role of TFs in the rise and adaptation of eukaryotes and the complexity arising within their successors.

Far more recently, a transposon-mediated shift in regulatory signaling in mammalian pregnancy has led to the endometrial stromal cell type [63] and rewiring of a stress response producing the decidual stromal cell type [64]. In humans, the development of our cognitive skills is possibly due to a delay of synaptic-function gene expression and corresponding synaptogenesis in the pre-frontal cortex, owing to transcription factors such as myocyte enhancer factor 2A (MEF2A) [65–68]. At the same time, evidence is accumulating that maturation occurred earlier in the closely-related Neanderthal, as it does in chimpanzee [69], which is supported by samples of jaw [70], tooth [71], and, most importantly, cranium [72]. Furthermore, development of globular brain structure was not present in modern humans at the time of divergence with Neanderthal and Denisovan lineages but has been a unique product of modern human development within the last 35,000–100,000 years [73] making it also particularly attractive for further analyses.

Modern human brain regulatory changes

The TF genes we have identified as differentially binding show a strong enrichment for homeobox genes, the primary regulators of development in humans and other multicellular organisms, compared to the complete set of JASPAR TFs, which are themselves very strongly enriched compared to all protein coding genes. A further six DB TFs are forkhead box genes, also known drivers of development. Similarly, the DB TFs are enriched for brain-specific expression compared to JASPAR TFs, which are again enriched compared to all human genes. Hierarchical clustering of TF expression revealed that DB TF genes associated with neural and immune function formed distinct

groups. Within the tissue axis, in the neural cluster, adult and fetal/newborn neural tissues cluster separately in several instances (Figure 1). From these discoveries in a broader set of data, it became important to focus more closely on the expression of these DB TFs in developing neural tissues specifically. Analysis of RNA-Seq data from whole brain subtissues further revealed differences in expression of these genes in a chrono, and likely developmentally, dependent manner. Specifically, expression of DB TFs in all tissues for the two groups occurring within 8 weeks post conception to 4 years post birth were largely segregated from all tissues for the group defined by 8 years to 40 years of age (Figure 2). Additionally, several modules of DB TFs identified by cluster analysis appear to show a cyclic pattern of expression during fetal development which then stabilizes post birth (Figure 2A).

DB TFs co-occur as DEGs in single-cell RNA-Seq of glutamatergic cells of L4 human cortex

Recent proliferation of single-cell transcriptomics has begun to clarify the direct role of transcription factors in cell type determination and identity (Arendt et al. 2019). To determine in which cell types DB TFs are expressed in adult brain, single-nucleus data for 49,417 cortical cells was examined. A subset of the DB TFs were annotated as commonly co-occuring DEGs: CUX1, CUX2, ESRRG, FOXP1, FOXP2, MEF2C, POU6F2, PRRX1, and RORA. Interestingly, we observed that these DB TF DEGs were primarily found co-occurring in glutamatergic cells of the L4 cortex. Within this set of nine genes, six of them (CUX1, CUX2, FOXP1, FOXP2, MEF2C, and RORA) are cataloged in the SFARI Gene knowledgebase [74] (sfari.org; vers. 3.0) as autism spectrum disorder (ASD) candidate genes. Likewise, six (CUX1, FOXP1, FOXP2, MEF2C, POU6F2, and RORA) are present in the genome-wide association studies (GWAS) catalog [75] (www.ebi.ac.uk/gwas) for the 'schizophrenia' (MONDO_0005090) term. Supporting this connection, a recent large-scale study on

schizophrenia genetics in 74,776 individuals with schizophrenia (and 101,023 control individuals) has shown enrichment for genes which express highly in glutamatergic neurons of both human and mouse cerebral cortex [76]. Similarly, a recent large-scale study on ASD genetics in 11,986 individuals with ASD (and 23,598 control individuals) identified 102 ASD risk genes, and showed that fetal-development excitatory neurons expressed the greatest number of these among all cell types identified [77].

Co-occurring DB TF DEGs associate with HARs and neuropsychiatric disease

Human-accelerated regions (HARs) are locations in the genome where the rate of evolutionary change has accelerated since divergence with chimpanzee. A study of 2,737 HARs has shown that they are enriched for TFBSs generally, and for TFs associated with neural development specifically (Doan et al. 2016). While the DB TFs were not enriched for HAR association compared to JASPAR TFs (OR 0.98; p-value 0.57), the JASPAR TFs are compared to all genes (OR 3.27; p-value 6.73 x 10^{-26}). Using the HARs cataloged by [57] we identified 28 of the DB TFs which are either in close proximity to a HAR or directly interact with one through chromatin conformation changes. Additionally, all but one (CUX2) of the co-occurring DEGs DB TFs (CUX1, ESRRG, FOXP1, FOXP2, MEF2C, POU6F2, PRRX1, and RORA) identified in the L4 excitatory neurons of cerebral cortex (Allen Brain Atlas scRNA-Seq data) contain HAR regions or HAR interactivity. Several of these contain or associate with HAR-variants that have been noted to cause neurological disease. POU6F2 possesses an intronic HAR where a mutant allele (GRCH38:chr7:39,033,595) is associated with ASD (Doan et al. 2016). The promoter of the CUX1 gene has been determined by ChIA-Pet analysis of chromatin to interact with a HAR, located ~200kb away; unrelated individuals with intellectual disability (ID) (IQ<40) and ASD have been identified with a homozygous mutation (GRCh38:chr7:101,606,361) in this HAR. Luciferase reporter assay indicates that when the mutant

version of this HAR interacts with the promoter of the CUX1 gene, its expression is increased by three-fold, while cultured differentiated neurons with enhanced CUX1 expression produced an increased synaptic spine density. Similarly, another HAR (GRCh38:chr5:88,480,873) is shown to interact with the promoter of MEF2C, and a mutation in this HAR creates a putative MEF2A binding site, reducing expression by ~50%. Mutations in the MEF2C gene are associated with autism [78,79], mental retardation [78–80], schizophrenia [81,82], epilepsy [78,79], and speech abnormalities [78]. Additionally, binding motifs for several of the DB TFs identified in our study were found to be enriched in HAR regions; POU6F1 and POU2F1 in ultra-conserved HARs, and HNF1A in all HARs [57]. After these intriguing results, further direct analysis of HARs using the TFBSFootprinter tool is merited.

Importantly, one of the identified DB TFs in our data is the FOXP2 TF gene, which is the first noted "speech gene" [83]. It has been shown that for the early postnatal period of mouse, FOXP2 is a negative regulator of MEF2C, and that the likely result is promotion of synaptogenesis of cortical striatum [84]. In addition to this interaction with MEF2C, there is a POU3F2 (which our results show is also a DB TF) binding site within the FOXP2 gene which has been shown to affect FOXP2 expression and which is associated with a selective sweep occurring since the divergence of humans and Neanderthal [85,86]. Because of its significantly documented role in human evolution, further analysis is warranted regarding regulation of FOXP2 in other relevant areas (e.g., introns, 3' UTR, enhancers, etc.) and for genome wide binding sites of FOXP2 itself.

DB TF target genes are expressed in oligodendrocytes, testis/sperm, maternal-fetal interface, and mitochondria

There were 74 genes identified as DB TF target genes. Of these, all but 8 have specific expression in reproductive (30), brain and retina (28), immune (18), or mitochondria (16) cells or tissues. Within

the 30 DB TF target genes with reproductive cell- and tissue-specific expression, the majority (19) are expressed in male tissues (testis, early and late spermatids, spermatocytes, spermatogonia). A total of 12 were expressed in female-specific cells and tissues, and in many cases specifically those involved in the maternal-fetal interface (cervix, granulosa cells, cytotrophoblasts, extravillous trophoblasts, syncytiotrophoblasts, and endometrial ciliated cells). Within the 28 DB TF target genes with brain- and retina-specific expression, a total of 16 have expression in oligodendrocytes, 12 in inhibitory neurons, and 11 in excitatory neurons.

The DB TF target gene with the greatest number of differential binding events in its promoter region is FARS2, a member of the mitochondrial aminoacyl-tRNA synthetases (mt-aaRSs). scRNA-Seq data from the Protein Atlas indicates FARS2 has enhanced expression in excitatory neurons, oligodendrocyte precursor cells, inhibitory neurons, oligodendrocytes, astrocytes, microglial cells, late spermatids, and early spermatids. mt-aaRSs charge their cognate mt-tRNA with the appropriate amino acid, and mutations in these genes cause a variety of diseases but most predominantly those which affect the central nervous system [87], perhaps due to delayed myelination, demyelination, or both [88]. In addition to their canonical roles, the mt-aaRSs are hypothesized to have functions in monitoring of amino acid levels, as sensors for the mitochondrial environment, and transcriptional regulation [87] and through the addition of new protein domains have been associated with neural development and immune response, among others, as reviewed by [89]. The FARS2 gene produces the mt-PheRS protein (phenylalanyl-tRNA synthetase) which is mitochondria-locating and responsible for attaching phenylalanine to its corresponding mt-tRNA for mitochondrial protein translation [90]. Intragenic variants in the FARS2 gene have been linked to two primary clinical manifestations, early-onset epileptic mitochondrial encephalopathy and spastic paraplegia, and for patients in both groups symptoms can also include intellectual disability

or developmental delay [90]. Deletions within FARS2 and reduced expression levels have also been associated with schizophrenia [91].

Importantly, FARS2 shares a bidirectional promoter with another mitochondrial gene, LYRM4 (LYR Motif-Containing Protein 4), which has a TSS just ~400 bp away. Together with ISCU, NFS1, and FXN, the ISD11 protein of the LYRM4 gene is involved in formation of iron–sulfur (Fe–S) clusters [92,93], essential cofactors in many basic biological processes such as formation of respiratory chain complexes I, II, III [94] and subsequently oxidative phosphorylation [95]. Mutations in LYRM4 have been noted to cause deficits in oxidation phosphorylation reactions [96], which are critical in neuronal development and schizophrenia as reviewed in [97]. Supporting this connection, polymorphisms occurring in the FARS2-LYRM4 bidirectional promoter region have been shown to be associated with cognitive deficit and schizophrenia [98]. Likewise, a 900kb microdeletion in this region (encompassing RPP40, PPP1R3G, LYRM4, and part of FARS2 and CDYL genes) has been shown to produce gyral pattern anomaly, intellectual disability and speech and language disorder [99].

A comparison of metabolite levels in pre-frontal cortex, visual cortex, cerebellum, kidney, and muscle between human and other primates (chimpanzee and macaque), showed 'aminoacyl-tRNA biosynthesis' as the most significant enriched pathway for metabolites having higher levels in human; for all three brain regions, but not muscle or kidney [100]. Likewise, the 'Phenylalanine, tyrosine and tryptophan biosynthesis' pathway was enriched for all three brain regions. This observation fits with the ongoing hypothesis that there exists a requirement for coevolution of nuclear and mitochondrial genes coding for mitochondrial proteins, known as mitonuclear compensatory coevolution, as reviewed in [101]; and that this coevolution is a driver of speciation [102,103]. In the mitonuclear compensatory coevolution hypothesis, nuclear-encoded genes

coding for aminoacyl tRNA synthetases (like FARS2) and OXPHOS complex components (like LYRM4) are expected, and in some cases have been shown, to evolve more rapidly in response to high rates of evolutionary change in mitochondrially-encoded genes, with which their protein products subsequently directly interact [101]. Another consideration is how to regulate translation of OXPHOS complex components when the corresponding genes exist in both mitochondrial and nuclear genomes [104,105]. The high number of observed DB TFBSs in the bidirectional promoter of the nuclear FARS2 and LYRM4 genes thus potentially allow more well-regulated coordination of cross-compartment OXPHOS component gene expression, and may even contribute to the effects of both mitonuclear compensation and speciation.

Taken all together, the LYRM4-FARS2 locus is potentially of great interest in the divergence of modern human with other hominins and hominids. It is possible to imagine that differences in neuronal development and function, and even diet, between MH and Neanderthal could require corresponding differences in mitochondrial activity, metabolism, and amino-acid sensing.

Lots of work left to do

The divergence of modern human and Neanderthal is estimated to have occurred between 400,000 and 800,000 years ago [106–108]. Since then both subspecies have experienced unique evolutionary, social, and cultural paths, many of which overlapped and intertwined. What appears to be both divergent and unifying in the hominin lineage is significant change to the brain leading to abilities in cognition, social function, language and creativity. In the case of humans, and likely for Neanderthal as well, many of the same genes which are at the core of our novel capabilities are the same which are commonly the root of our maladies. Autism and schizophrenia, among other neuropsychiatric disorders, appear to be maintained in humans at a greater rate than would be expected if they were strictly deleterious: the global prevalence of ASD is about 1% [109] and

0.28% for schizophrenia [110]. Likewise, as the study of these diseases continues, the direction of understanding tends towards a continuum of effects rather than discrete on/off state. A study by Linscott et al. reports prevalence of psychotic experiences in the general population at 7.2% [111]. Our results have shown that DB TFs are enriched for development and brain, and as studies on HAR regions have shown mutations in regulatory regions alone are sufficient to cause significant neuropsychiatric effects [57]. As a result, there appears to be significant opportunities to analyze gene regulation for understanding not only the evolution of human cognitive abilities but also the neuropsychiatric disorders which have accompanied it.

Limitations

We have not included analyses of 3' UTR, introns, or enhancers. The single-most expressed transcript for each protein-coding gene was chosen for analysis, while MH vs. Neanderthal SNPs may have occurred more closely to TSSs of lower-expressed transcripts and thereby have higher and perhaps more significant scoring. Only transcripts for protein-coding genes were analyzed, despite the existence of numerous other transcribed sequence types. The SNP dataset we used for analysis is based on multiple Neanderthal individuals and has a lower sequencing coverage than newer datasets. The inclusion of multiple individuals is useful for comparing modern humans to Neanderthals as a group, but can only provide a more general comparison. Using a higher coverage dataset would allow greater assurance that SNPs under inquiry are legitimate. We chose to perform the TFBSFootprinter analysis using all transcription-relevant features available, in the future we plan to expand testing and assessment of empirical datasets and incorporate an option to use the combination of features which is proven best for each individual TF.

TFBSFootprinter

The TFBSFootprinter tool incorporates 7 different transcription-relevant empirical data features in the prediction of TFBSs. It can take as input any Ensembl transcript ID from any of 125 vertebrate species available in the Ensembl database. Starting with a list of Ensembl transcript ids for a target species (e.g., Homo sapiens) TFBSFootprinter will retrieve from the Ensembl REST server, a user-defined region of DNA sequence surrounding each transcription start site (TSS). The sequence is then scored using up to 575 JASPAR TFBS profiles, or a more limited set as defined by the user. User-defined p-values may be used to filter results; and the corresponding score thresholds have been determined by scoring each JASPAR TFBS profile on the complete human genome. Each putative TFBS is then additionally scored based on proximity/overlap with TSS, TFBS metaclusters, open chromatin, and eQTLs which affect expression levels of the proximal gene, as well as conservation of sequence, correlation of expression with proximal (target) gene, and CpG content. Additional transcription-relevant data can be added easily in the future as is appropriate.

We believe that TFBSFootprinter provides an excellent way to predict TFBSs, thus easily supplementing current investigations into gene function or providing a means to perform larger-scale analyses of groups of related target genes. The ability to identify conserved binding sites in a large number of species particularly widens its applicability to researchers studying various vertebrates. After completion of analysis a publication-ready figure depicting the top scoring TFBS candidates is produced. Additionally, a number of tables (.csv) and JavaScript Object Notation (.json) files presenting various aspects of the results are output. Primary among these is a list of computational predictions in the target species which are supported by empirical data, sorted by a sum of the combined log likelihood scores (the combined affinity score). Importantly, scoring of non-human species becomes limited by the availability of external data for that species;

at this time the only data available for non-human species are sequence conservation, CpG, and JASPAR motifs.

Methods

Identification of the canonical promoter

Analysis of differences in TF binding affinity between MH and Neanderthal promoters requires defining the promoter search space, and accordingly the target promoter boundaries. To this end, the complete set of ChIP-Seq data (15,982 experiments; 1,391 TFs; and 1,039 cell types and tissues) from the GTRD database (gtrd.biouml.org; version 20.06) was downloaded for review [112]. The single most highly expressed transcript from each Ensembl gene object was identified, a total of 40,490 transcripts, as described by transcript-level expression data from GTEx dataset (v8). Subsequently, all ChIP-Seq peaks with a fold-enrichment ≥10 (58,976,174 peaks) were mapped to any overlap of the region ±40,000 bp relative to the TSS of all most highly expressed transcripts (Figure 7). From this analysis the sequence ±2,500 bp relative to transcript TSS was defined as the region of interest for identifying promoter region MH-Neanderthal SNPs.

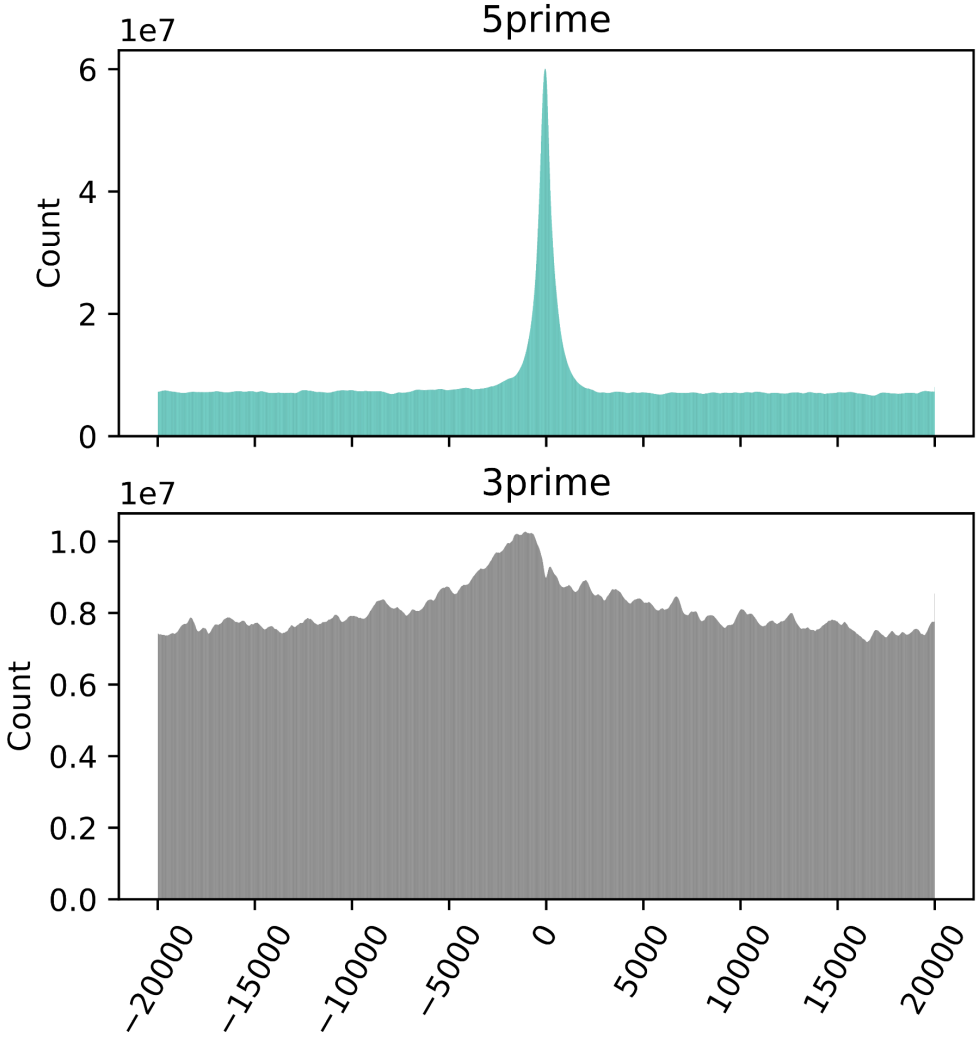

Figure 7. Identification of regulatory region occupancy by TFs. Mapping of 58,976,174 ChIP-Seq peaks (fold-enrichment ≥10) cataloged in the GTRD database [112] to the 40,490 most highly expressed transcripts (max 1 per gene), as defined by transcript-level expression data from the GTEx project.

Analysis of modern human vs. Neanderthal genetic variation

The Neanderthal Genome Project has cataloged a total of 388,388 SNPs in comparing modern human vs. Neanderthal genomes [1]. These were further reduced to 21,990 SNPs in the proximal promoters (-2,500 to +2,500 nts relative to the TSS) of those transcripts which are defined by Ensembl as "protein-coding" (Figure 8C). Using the TFBSFootprinter tool, a 50bp region centered on each SNP was analyzed for binding of 575 TFs, for both the modern human version and the

Neanderthal variant. TFBSFootprinter automatically retrieves the human sequences at a target region, and custom Python scripts were used to modify these sequences for the Neanderthal variant. All TFBS predictions which overlap the target SNP position were kept. The complete result set was then reduced based on the combined affinity score p-value using a Benjamini-Hochberg derived critical p-value corresponding to a false discovery rate cutoff of 0.01 to address multiple testing. For each putative TFBS meeting the cutoff in either subspecies, the corresponding matched pair of PWM scores was kept. To identify statistically different scoring for TFs between subspecies: for each TF, using the compiled matched scores, the Wilcoxon rank statistical test was performed using the SciPy stats Python library [113] and subsequent results filtered using a Benjamini-Hochberg derived critical p-value corresponding to a false discovery rate cutoff of 0.10 to address multiple testing.

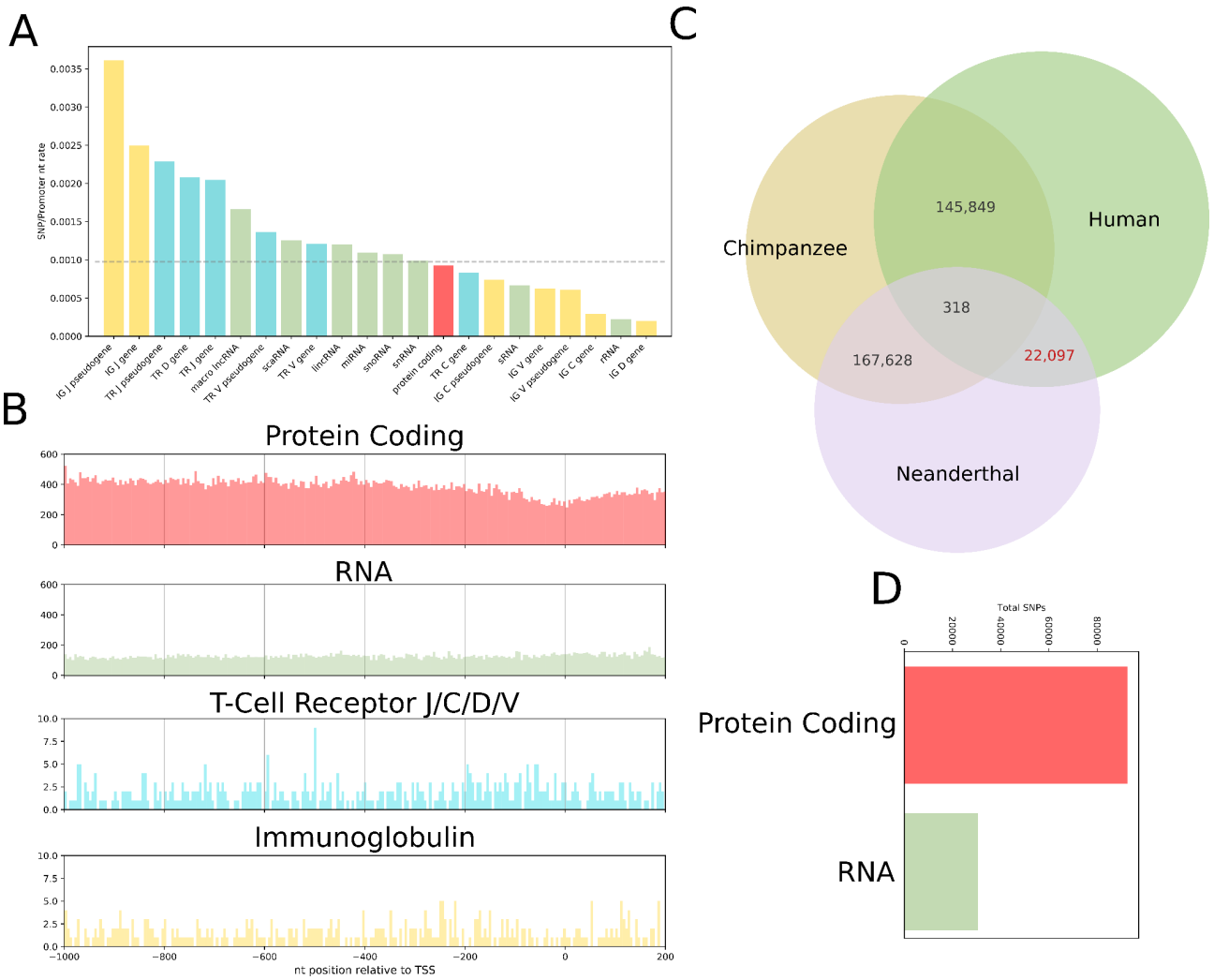

Figure 8. Modern human and Neanderthal variants used in analysis. (**A**) Incidence rate of SNPs/nt in human promoters, of the various transcript classes, as defined by Ensembl. Bars are colored by transcript class, correspondent with panel B. (**B**) Number of observed modern human/Neanderthal SNPs at each nt location in the sets of promoters of various transcript types. A visible depletion of SNPs is observed for the promoters of protein-coding transcripts corresponding with increasing proximity to the TSS (nt position 0). (**C**) Counts of total SNPs in promoters of protein-coding transcripts, between the three species cataloged in the Neanderthal Genome Project. (**D**) Comparison of counts of human vs. chimp SNPs in promoters of protein-coding vs RNA transcripts. IG J, immunoglobulin joining chain; TR J, T-cell receptor joining chain; TR D, T-cell receptor diversity chain; TR V, T-cell receptor variable chain; IG C, immunoglobulin constant chain; IG V, immunoglobulin variable chain; IG D, immunoglobulin diversity chain.

A total of 108 TF models were identified as scoring differently across human vs. Neanderthal SNP locations. For each of these we extracted RNA-Seq data from the FANTOM data set (across all CAGE peaks associated with that TF) and kept the data for the 100 tissues with the highest aggregate expression across all of the target TF genes. In the case of hetero-dimer JASPAR TF

models (e.g., FOS::JUN, NR1H3::RXRA, and POU5F1::SOX2) the expression of each TF component gene was used. Expression was extracted as TPM values and was normalized by $\log_2$ transformation. From the subsequent normalized expression data values, hierarchical clustering was performed and visualized using SciPy, Matplotlib, and Seaborn Python libraries [113–115].

The results of hierarchical clustering revealed a cluster of TF genes which showed unique expression in neural and immune tissues. These gene sets were used to perform PANTHER-based gene ontology enrichment analysis (using the www.geneontology.org web server) [51,52], with the default statistical settings using the Fisher's Exact test method and a FDR threshold of $p<0.05$.

Analysis of brain expression of differentially binding TFs

Expression data in the form of reads per kilobase per million reads (RPKM) were extracted for 26 unique tissues across 31 timepoints (from 8 weeks post-conception to 40 years after birth) from the Allen Brain Atlas (brainspan.org). RPKM values were then converted to TPM and $\log_2$ transformed. Ages were grouped into three phases of growth for simplicity of analysis and interpretation: pcw (8 to 37 weeks post conception), early (4 months to 4 years), and late (8 years to 40 years). Correspondingly, log-transformed TPM values were grouped and averaged and used to perform clustermap analysis to identify grouping tissues at time phases with similar expression profiles.

Analysis of brain sample scRNA-Seq

The Allen Brain Atlas has performed single-nucleus RNA-Seq analysis of 49,417 nuclei derived from 8 brain cortex regions within the middle temporal gyrus (MTG), anterior cingulate gyrus (CgG), primary visual cortex (V1C), primary motor cortex (M1C), primary somatosensory cortex (S1C) and primary auditory cortex (A1C) [55]. These data were downloaded as a count matrix along with a table of the associated metadata

(https://portal.brain-map.org/atlases-and-data/rnaseq/human-multiple-cortical-areas-smart-seq) and loaded into Python library SCANPY [116]. Using a modified workflow described previously in [117], samples were filtered by Gaussian fit of read count (300,000<x<3,500,000), expressed gene count (2,000<x), and number of cells in which a gene is expressed (>50), resulting in a final count of 46,959 cells and 42,185 genes for further analysis. Using SCANPY, counts were normalized by cell ('pp.normalize_total'; target_sum=1,000,000), log transformed ('pp.log1p'), highly variable genes genes identified ('pp.highly_variable_genes'; flavor='seurat_v3'; n_top_genes=5,000; layer='counts'), principal component analysis ('pp.pca'; n_comps = 15; svd_solver='arpack'), and k-nearest neighbors ('pp.neighbors'; n_neighbors=15). Expression relationships between cells were graphically visualized with the Python implementation [118] of the ForceAtlas2 [119] graph layout algorithm as called in Scapny ('tl.draw_graph' and 'pl.draw_graph').

Annotation data regarding brain region, cortical layer, and GABAergic/glutamatergic/non-neuronal cell type features were extracted from Allen Brain Atlas sample data for mapping onto derived cell type clusters. The top 100 marker genes for each cell type+cortical layer (CT/CL) cluster (e.g., excitatory L4) were identified as those with higher expression unique to each cluster by Welch t-test in SCANPY ('tl.rank_genes_groups'). Expression of the DB TF list genes which were identified as a marker gene in a CT/CL cluster were mapped onto cluster figures.

Gene ontology analysis of target cluster marker genes was performed using the Protein Analysis Through Evolutionary Relationships (PANTHER) tool at the geneontology.org web server [52]. Ontological term overabundance among cluster marker gene lists were established by Fisher's exact test and results were filtered by FDR<0.05; analyses were performed for biological process, molecular function, and cellular component terms. Disease term gene ontology analysis was performed using Enrichr [56,120] based on ontology compiled by DisGeNET [121].

TFBSFootprinter Methodology and Scoring

A computational pipeline was created in Python to allow for automated vertebrate promoter sequence retrieval from the Ensembl database (Ensembl version 94 was used in this analysis). The user-defined target sequence is then analyzed using 575 different transcription factor position weight matrices (PWMs), or a user-defined subset, derived from PFMs taken from the JASPAR database. Each TFBS prediction results in a log-likelihood score indicating the likelihood of a particular TF binding the DNA at that location. After this initial step, seven additional gene transcription related features are assessed for each TFBS prediction, each of which generate their own log-likelihood score based on proximity or overlap with these features. The features which may be considered for each TFBS prediction are: vertebrate sequence conservation (GERP), proximity to CAGE peaks (FANTOM5), correlation of expression between target gene and TF predicted to bind promoter (FANTOM5), overlap with ChIP-Seq TF metaclusters (GTRD), overlap with ATAC-Seq peaks (ENCODE), eQTLs (GTEx), and observed/expected CpG ratio (Ensembl). A summation of these scores, for each putative TFBS, then equals a value which we describe as the 'combined affinity score'. In this way the model's parameters are significantly more empirically flexible and therefore robust, and ultimately generate a more complete picture of a binding site instead of just computational prediction of binding affinity to a static set of nucleotides.

Ensembl sequence retrieval

The Ensembl Representational State Transfer (REST) server application programming interface (API) (Yates et al. 2014) is used by TFBSFootprinter for automated retrieval of user-defined DNA sequence near the transcription start site of an established Ensembl transcript ID. Annotations for the transcript, and Ensembl-defined regulatory regions (e.g., 'promoter flanking region') are also retrieved and mapped in the final output figure.

PWMs

A total of 575 transcription factor PFMs retrieved from the 2018 JASPAR database [22] (http://jaspar.genereg.net/) are used to create PWMs (Eq 1), as described by (Nishida, Frith, and Nakai 2009):

$$weight_{binding} = \sum_{i=1}^{N} \frac{\frac{(a_i + b/4)}{(S+b)}}{\frac{(n_{nuc} + b)}{(l_{bg} + b)}} \quad \textbf{(1)}$$

N is the set of nucleotides in the currently scanned sequence; a_i is the number of instances of nucleotide *a* at position i; *b* is a pseudocount set to 0.8 as per [122]; *S* is the number of sequences describing the motif; n_{nuc} is the count of the nucleotide in the background sequence; and l_{bg} is the length of the background sequence. The background frequencies for each nucleotide were set to match those of the human genome as determined previously [123].

Without a score threshold, scoring of a 1,000 nt promoter with all 575 JASPAR TF profiles will produce ~1,150,000 predictions. In order to address this issue, each of the 575 JASPAR PWMs was used to score the complete human genome (a total of 3,375,096,897,466 TF-DNA binding calculations). Scores for each TF were then used to generate a distribution which allowed pairing scores with p-values (from 10^{-1} to 10^{-5}, or smaller when possible, depending on TF profile length). As a result, an appropriate score threshold can be chosen at the discretion of the user. Computation was performed using the supercomputing resources of CSC – IT Center for Science Ltd. (a non-profit owned by the state of Finland and Finnish higher education institutions).

CAGE peak locations and Spearman correlation of expression values

Cap analysis of gene expression (CAGE) uses sequencing of cDNA generated from RNA to both determine TSSs and quantify their expression levels. The FANTOM project has performed CAGE

across the human genome [124] and the results are freely available for download (http://fantom.gsc.riken.jp/data/). Using the genomic locations of FANTOM CAGE peaks, the distances from all human nucleotides to the nearest CAGE peak were calculated. The distribution of these distances was used to generate a log-likehood score for all observed distances. The CAGE peak locations and distance/log-likelihood score pairings are then used during de novo prediction of TFBSs (Eq 2).

$$weight_{CAGE\ distances} = \sum_{i=0}^{N} P(human\ genome) * \frac{p_i}{p_{total}} \quad \textbf{(2)}$$

Where N is the number of all CAGE peaks associated with the target gene; d_i is the distance to current CAGE peak; p_i is the number of peak counts of current CAGE peak; and p_{total} is the total peak count for this gene.

Expression data for CAGE peaks associated with the 575 JASPAR TF genes was then combined with expression data for all CAGE peaks to perform a total of 386,652,770 Spearman correlation analyses using the 'spearmanr' function from the SciPy Stats module [113]. Bonferroni correction was performed to account for multiple testing. Due to the size of the analysis, a cutoff correlation magnitude value of 0.3 was used, and all lesser values (-0.3<x<0.3) and correlation pairs were discarded. A distribution was generated from the resulting correlation data which was used to generate log-likelihood scores for each possible correlation value (Eq 3). The CAGE peak expression correlations/log-likelihood score pairings are then used during de novo prediction of TFBSs. Computation was performed using the supercomputing resources of the CSC – IT Center for Science Ltd.

$$weight_{expression\ correlation} P\left(c_{all}\right) \quad \textbf{(3)}$$

Where $c_{current}$ is the Spearman correlation between the expression of the target gene and the expression of the TF corresponding to the putative TFBS; and c_{all} is the distribution of all Spearman correlations between JASPAR TF genes and all genes.

Experimental TFBSs compiled by the GTRD

Experimental data on TFBS binding (derived from ChIP-Seq, HT-SELEX, and PBMs) is one of the most direct ways to locate potentially functional TFBSs. While binding alone is not indicative of function, clusters of binding sites have been shown to imply functionality [27,39]. The GTRD project (gtrd.biouml.org) is the largest comprehensive collection of uniformly processed human and mouse ChIP-Seq peaks and has compiled data from 8,828 experiments extracted from the gene expression omnibus (GEO), sequence read archive (SRA), and the encyclopedia of DNA elements (ENCODE) databases [112]. One of the outputs of the performed analyses are reads that have been grouped to identify metaclusters, places where TF binding events cluster together in the human genome. We retrieved the metacluster data (28,524,954 peaks) from the GTRD database (version 18.0) and subsequently mapped the number of overlapped metaclusters for all human nucleotides. The distribution of these overlaps was used to generate a log-likehood score for all observed distances. The metacluster locations and distance/log-likelihood score pairings are then used during de novo prediction of TFBSs (Eq 4).

$$weight_{metaclusters} P\left(D_{human\ genome}\right) \tag{4}$$

Where $n_{overlap}$ is the number of metaclusters overlapped by the current putative TFBS; and $D_{human\ genome}$ is the distribution of the number of overlapping metaclusters for every location in the human genome.

ATAC-Seq peaks

Assay for Transposase-Accessible Chromatin using sequencing (ATAC-Seq) is an experimental method for revealing the location of open chromatin [125]. These locations are indicative of genomic regions which, due to their unpacked nature, may allow TFs to bind to DNA and subsequently influence transcription. Open chromatin regions have been shown to be useful in the prediction of TFBSs [126]. We retrieved and compiled data from 135 ATAC-Seq experiments stored in the ENCODE project database (www.encodeproject.org) and mapped the distance from all human nucleotides to the nearest ATAC-Seq peak, and the distribution of these distances was used to generate a log-likehood score for all observed distances. The ATAC-Seq peak locations and distance/log-likelihood score pairings are then used during de novo prediction of TFBSs (Eq 5).

$$weight_{ATAC-Seq\ distances} = \sum_{i=0}^{N} P\left(D_{human\ genome}\right) \tag{5}$$

Where N is the number of ATAC-Seq peaks within the current target region; d_i is the distance to the current ATAC-Seq peak; and $D_{human\ genome}$ is the distribution of the distances from all human nucleotides to the nearest ATAC-Seq peak.

eQTLs

The genome tissue expression (GTEX) project (gtexportal.org) has performed expression quantitative trait loci (eQTL) analysis on 10,294 samples from 48 tissues from 620 persons (version 7) [45]. This analysis has identified 7,621,511 variant locations in the genome, usually 1–5 nt, that affect gene expression. eQTL data was extracted from the GTEX database and used to construct a distribution of the magnitude of effect on gene expression, which was then used to generate log-likelihood scores (Eq 6). Next, we generated a second distribution of the distance from each gene to its variants; the distance was limited to 1,000,000 bp from either end of the transcript as

this is the search area over which GTEx scans for variants affecting expression of each gene. The variant locations, magnitude of effect/log-likelihood score pairings, and distance/log-likelihood score pairings are then used during de novo prediction of TFBSs (Eq 7).

$$weight_{eQTL\ magnitude} = \sum_{i=0}^{N} P\left(D_{human\ genome}\right) \quad \textbf{(6)}$$

Where N is the number of eQTLs overlapping the current putative TFBS; m_i is the magnitude of effect of an eQTL overlapping the current putative TFBS; and $D_{human\ genome}$ is the distribution of all eQTL magnitudes in all human nucleotides.

$$weight_{eQTL\ overlap} = \sum_{i=0}^{N} P\left(D_{human\ genome}\right) \quad \textbf{(7)}$$

Where N is the number of eQTLs overlapping the current putative TFBS; l_{tf} is the length of the current putative TFBS (nucleotides); n_v is the number of variants associated with the target gene; l_g is the length of the current gene (nucleotides); l_t is the length of the target transcript; and $D_{human\ genome}$ is the distribution of all eQTL overlaps for nucleotide windows the same length as the current TFBS, in the human genome.

CpG islands

Due to the fact that methylation of DNA acts as a repressor of transcription, active promoters tend to be un-methylated. When methylated, the cytosine in a CpG dinucleotide can deaminate to thymine. Therefore, a CpG ratio close to what would be expected by chance is often indicative of an active promoter region [127,128]. Subsequently, CpG ratios (observed/expected) across a 200 nt window were computed for all human nucleotides. A distribution of these ratios was generated

and used to generate log-likelihood scores for each possible ratio. CpG ratio/log-likelihood score pairings are then used during de novo prediction of TFBSs.

$$weight_{CpG} P\left(D_{genome}\right) \quad \textbf{(8)}$$

Where $r_{obs/exp}$ is the ratio of observed to expected CpG dinucleotides in a 200 nt window centered on the current putative TFBS; and D_{genome} is the distribution of $r_{obs/exp}$ across all nucleotide locations in the target genome.

Conservation of vertebrate DNA

Conservation of DNA across species boundaries is evolutionarily costly and thus implies function. While differential regulation of gene transcription is what accounts for much of the variation among species [129], there is conservation of regulatory elements across closely related species groups, such as among primates [130]. The Ensembl EPO (Enredo, Pecan, Ortheus) pipeline has created whole genome multiple sequence alignments for distinct clades of vertebrates [131]. As of Ensembl release 94, whole-genome alignments exist for mammals (70 species), fish (48 species), amniotes (32 species), and sauropsids (7 species). The number of available species is dependent on the current Ensembl release and is continually growing. Conservation of sequence analysis has been performed by Ensembl to identify constrained elements for each species in each species group using the genomic evolutionary rate profiling (GERP) tool [132]. For each of the vertebrate species of Ensembl release 94, we have calculated the distance from all nucleotides in the associated species genome to the nearest GERP constrained element; and generated distributions of distances which were used to calculate log-likelihood scores for each distance. GERP element distance/log-likelihood score pairings, for each species, are then used during de novo prediction of TFBSs in the relevant species.

$$weight_{conservation} P\left(D_{genome}\right)$$

(9)

Where d_i is the distance between the current putative TFBS and the nearest conserved element in an alignment of 70 mammalian genomes (GERP); and D_{genome} is the distribution of distances between all nucleotides in the target genome and the nearest GERP conserved element.

Combined affinity score

A summation of the weight scores from each experimental dataset is then performed for each putative TFBS and is represented as the 'combined affinity score'. For analysis of human sequences this is represented by Eq 10. Due to the limitations of available experimental data for non-human species, currently, for vertebrates the combined affinity score is described by Eq 11. A complete scoring of ~80,000+ transcript's promoter regions (1,000 nt) was used to generate p-values for combined affinity scoring; computation was performed using the supercomputing resources of CSC – IT Center for Science Ltd.

$$Combined\ Affinity_{human} = Eq_1 + Eq_2 + Eq_3 + Eq_4 + Eq_5 + Eq_6 + Eq_7 + Eq_8 + Eq_9$$

(10)

$$Combined\ Affinity_{non-human} = Eq_1 + Eq_8 + Eq_9$$

(11)

Benchmarking TFBSFootprinter

In order to robustly assess the ability of PWM, TFBSFootprinter, and DeepBind models to identify functional human transcription factor binding sites, several approaches were taken for benchmarking. These included using ChIP-Seq peaks from the GTRD database (compiled and uniformly processed from ENCODE, GEO, and SRA databases) as well as experimentally validated

functional regulatory binding sites to serve as true positives. True negatives were also defined using several approaches in order to mitigate biases inherent in each approach.

ChIP-Seq peaks for 1,392 transcription factors from 15,982 experiments uniformly processed with MACS2 [133] were downloaded from the GTRD database. Using the full complement of Ensembl (release 100) protein-coding transcripts, the ChIP-Seq peaks were reduced to those which occur within the window -1000 to +200 bp relative to an Ensembl transcript TSS. In the first benchmark, for each JASPAR TF with an identically named match in this subset of GTRD peaks, the top 100 peaks which had a fold-enrichment of ≥50x were kept as true positives and matched locations 1,000 bp upstream were used as true negatives. In the second benchmark, for each JASPAR TF with an identically named match in this subset of GTRD peaks, the top 100 peaks which had a fold-enrichment of ≥50x were kept as true positives and the bottom 100 peaks which had a fold-enrichment of ≤2x were kept as true negatives. Using this set of high and low-occupancy locations, a 50-bp window centered on the peak summit was defined for analysis with TFBSFootprinter and DeepBind.

Experimentally verified and curated TFBSs belonging to the annotated regulatory binding sites (ABS) [134], ORegAnno [135], and Pleiades promoter project [136] databases were retrieved as curated GFF files from the Pazar database [137]. The TFs associated with each experimental TFBS were identified using a combination of Python scripting and manual parsing, due to deprecated gene naming in some cases. From this data, 504 experimentally-validated binding sites affecting gene expression for 20 DeepBind TFs and 607 experimentally-validated binding sites affecting gene expression for 25 JASPAR TFs were selected. Selection criteria for the chosen TFs was that they have at least 10 experimentally-validated binding sites affecting gene expression. All target sites were converted from Hg19 to GRCh38 genomic coordinates using Ensembl REST. Subsequently, 50

bp sequences were retrieved centered on each experimentally-validated functional binding site in the human genome, to serve as true positives.

Paired with these true positives were true negatives from two different sources. In the first approach, for each true positive, 50 locations were chosen in random Ensembl transcripts, at the same distance from the TSS. In the second approach, the 50 locations were chosen within the promoter of the same Ensembl transcript, at least 50 nucleotides away from the true positive.

For each of the TFBSFootprinter, PWM, and DeepBind methods, the true positive locations were scored using that method and the model for the corresponding TF which had previously been experimentally verified as functional at that location. However, for the set of true negatives, each method scored all locations using all TF models.

Analyzing effect of feature combinations on TFBSFootprinter accuracy

Subsequently, for each method, and for each TF, the correlating true positive and true negative scores were used to generate receiver operator characteristic (ROC) curves using the ‘roc_curve’ module of the scikit-learn Python library [138].

Using the TFBSFootprinter tool, all 128 possible combinations of transcription-relevant features (PWM, CAGE, eQTL, metaclusters, ATAC-Seq, CpG, sequence conservation, expression correlation) which include PWM as a component were used in scoring of the true positives and true negatives. This allowed identification of the best possible feature-combination TFBSFootprinter model for each TF, which is described as 'TFBSFootprinter best by TF', as well as the TFBSFootprinter model which performed best on average across all TFs, named as 'TFBSFootprinter best overall'. In assessment of the DeepBind tool both available models, based on SELEX or ChIP-Seq data, were used. Using related T-test, comparisons of ROC scores were made between PWM and each of TFBSFootprinter best by TF, TFBSFootprinter overall best, TFBSFootprinter all features, and

DeepBind best by TF. Similarly, comparisons were made between DeepBind and each of TFBSFootprinter best by TF, TFBSFootprinter overall best, TFBSFootprinter.

Data availability

Code reproducing the analyses and figures in the manuscript are available on GitHub at https://github.com/thirtysix/TFBS_footprinting_manuscript. Results of benchmarking of TFBSFootprinter (https://osf.io/hzny6/) and MH vs. Neanderthal promoter analysis (https://osf.io/r2mtw/) are available as Open Science Foundation repositories.

Funding

This work was supported by the Finnish Cultural Foundation and Fimlab to HB, and Academy of Finland and Jane & Aatos Erkko Foundation to SP.

Acknowledgements

Heini Huhtala is acknowledged for assistance in statistical techniques and professor Matti Nykter and Payam Emami Khoonsari PhD are gratefully thanked for discussions on practical and theoretical concerns. The non-profit CSC – IT Center for Science Ltd, owned by the state of Finland and Finnish higher education institutions, is acknowledged for providing computational resources for analyses.

Conflicts of Interest

Authors declare they have no conflicts of interest.

Figure supplement legends

Supplementary Figure 1. Aggregate expression of DB TFs in 100 tissues. FANTOM5 RNA-Seq data was extracted as TPM values and the 100 tissues with highest aggregate expression of the differentially binding TF genes were selected for clustering (Figure 1), resulting in order of labeled tissues depicted.

Supplementary Figure 2. ROC Analysis in the identification of TFBS ChIP-Seq peaks – strong and distal binding. ROC Curves are presented for the best performing TFBSFootprinter model for each TF.

Supplementary Figure 3. ROC Analysis model performance in the identification of experimentally verified functional TFBSs – random Ensembl transcripts. ROC analysis was performed using experimentally verified functional TFBSs as annotated in the ORegAnno/Pleides/ABS datasets as true positives, where true negatives were random locations in other Ensembl transcripts at the same distance from the TSS as the associated true positive. All ROC curve analyses were performed on TFs which had at least 10 true positives and 50 true negatives per true positive were used for each analysis. Each true positive/negative segment analyzed was 50 nucleotides long, and the highest TFBS score for the relevant dataset(s) was used for each true positive/negative. (A) Barplot of the frequency of experimental data type in the top 20 performing TFBSFootprinter models. (B) Boxplot of ROC scores for TFBSFootprinter and DeepBind for 14 TFs (left). ROC scores were also calculated based on using individual experimental metrics to show how well each contributes to accuracy of the combined model. (C) ROC scores for each individual TF tested, for each primary TFBS prediction model under study. The best scoring model among all is named for each TF (right). TFBSFootprinter best by TF, based on using the highest ROC score achieved by some combination of experimental data models; TFBSFootprinter overall best, based on using the combination of experimental data models which had the best average ROC score across all TFs analyzed; DeepBind best by TF, based on using the higher ROC score of the SELEX or ChIP-Seq DeepBind models.

Supplementary Figure 4. ROC Analysis model performance in the identification of experimentally verified functional TFBSs – random location in same Ensembl transcript. ROC analysis was performed using experimentally verified functional TFBSs as annotated in the ORegAnno/Pleides/ABS datasets as true positives, where true negatives were random locations in the same Ensembl transcript apositive. All ROC curve analyses were performed on TFs which had at least 10 true positives and 50 true negatives per true positive were used for each analysis. Each true positive/negative segment analyzed was 50 nucleotides long, and the highest TFBS score for the relevant dataset(s) was used for each true positive/negative. (A) Barplot of the frequency of experimental data type in the top 20 performing TFBSFootprinter models. (B) Boxplot of ROC scores for TFBSFootprinter and DeepBind for 14 TFs (left). ROC scores were also calculated based on using individual experimental metrics to show how well each contributes to accuracy of the combined model. (C) ROC scores for each individual TF tested, for each primary TFBS prediction model under study. The best scoring model among all is named for each TF (right). TFBSFootprinter best by TF, based on using the highest ROC score achieved by some combination of experimental data models; TFBSFootprinter overall best, based on using the combination of experimental data models which had the best average ROC score across all TFs analyzed; DeepBind best by TF, based on using the higher ROC score of the SELEX or ChIP-Seq DeepBind models.

Supplementary Figure 5. ROC Analysis in the identification of TFBS ChIP-Seq peaks – strong and weak binding. ROC Curves are presented for the best performing TFBSFootprinter model for each TF.

References

1. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. RESEARCH ARTICLE A Draft Sequence of the Neandertal Genome. Science. 2010;328: 710–723. doi:10.1126/science.1188021

2. Castellano S, Parra G, Sanchez-Quinto FA, Racimo F, Kuhlwilm M, Kircher M, et al. Patterns of coding variation in the complete exomes of three Neandertals. Proceedings of the National Academy of Sciences. 2014;111: 6666–6671. doi:10.1073/pnas.1405138111

3. Prüfer K, Hajdinjak M, Vernot B, Skov L, Hsieh P, Peyrégne S, et al. A high-coverage Neandertal genome from Vindija Cave in Croatia. Science. 2017;358: 655–658. Available: http://science.sciencemag.org/

4. Hajdinjak M, Fu Q, Hubner A, Petr M, Mafessoni F, Grote S, et al. Reconstructing the genetic history of late Neanderthals. Nature. 2018;555: 652–656. doi:10.1038/nature26151

5. Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, et al. Genetic history of an archaic hominin group from Denisova cave in Siberia. Nature. 2010;468: 1053–1060. doi:10.1038/nature09710

6. Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, Mallick S, et al. A high-coverage genome sequence from an archaic Denisovan individual. Science. 2012;338: 222–226. doi:10.1126/science.1224344

7. Prufer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, et al. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature. 2014;505: 43–49. doi:10.1038/nature12886

8. Petr M, Paabo S, Kelso J, Vernot B. Limits of long-term selection against Neandertal introgression. Proc Natl Acad Sci U S A. 2019;116: 1639–1644. doi:10.1073/pnas.1814338116

9. Zuckerkandl E. Perspectives in Molecular Anthropology. In: Washburn S, editor. Classification and Human Evolution. New York: Routledge; 1964. p. 265. doi:10.4324/9781315081083

10. King MC, Wilson AC. Evolution at two levels in humans and chimpanzees. Science. 1975;188: 107–116. doi:10.1126/science.1090005

11. Chen H, Li C, Zhou Z, Liang H. Fast-Evolving Human-Specific Neural Enhancers Are Associated with Aging-Related Diseases. Cell Syst. 2018;6: 604–611 e4. doi:10.1016/j.cels.2018.04.002

12. Liang H, Lin YS, Li WH. Fast evolution of core promoters in primate genomes. Mol Biol Evol. 2008;25: 1239–1244. doi:10.1093/molbev/msn072

13. Castellanos M, Mothi N, Munoz V. Eukaryotic transcription factors can track and control their target genes using DNA antennas. Nat Commun. 2020;11: 540. doi:10.1038/s41467-019-14217-8

14. Stormo GD, Schneider TD, Gold L, Ehrenfeucht A. Use of the “Perceptron” algorithm to distinguish translational initiation sites in E. coli. Nucleic Acids Res. 1982;10: 2997–3011. doi:10.1093/nar/10.9.2997

15. Cheneby J, Gheorghe M, Artufel M, Mathelier A, Ballester B. ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. Nucleic Acids Res. 2018;46: D267–D275. doi:10.1093/nar/gkx1092

16. Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, et al. DNA-binding specificities of human transcription factors. Cell. 2013;152: 327–339. doi:10.1016/j.cell.2012.12.009

17. Ogawa N, Biggin MD. High-throughput SELEX determination of DNA sequences bound by transcription

factors in vitro. Methods Mol Biol. 2012;786: 51–63. doi:10.1007/978-1-61779-292-2_3

18. Berger MF, Philippakis AA, Qureshi AM, He FS, Estep PW 3rd, Bulyk ML. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat Biotechnol. 2006;24: 1429–1435. doi:10.1038/nbt1246

19. Siddharthan R. Dinucleotide weight matrices for predicting transcription factor binding sites: generalizing the position weight matrix. PLoS One. 2010;5: e9722. doi:10.1371/journal.pone.0009722

20. Zhao Y, Ruan S, Pandey M, Stormo GD. Improved models for transcription factor binding site identification using nonindependent interactions. Genetics. 2012;191: 781–790. doi:10.1534/genetics.112.138685

21. Mathelier A, Wasserman WW. The next generation of transcription factor binding site prediction. PLoS Comput Biol. 2013;9: e1003214. doi:10.1371/journal.pcbi.1003214

22. Khan A, Fornes O, Stigliani A, Gheorghe M, Castro-Mondragon JA, van der Lee R, et al. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Res. 2018;46: D260–D266. doi:10.1093/nar/gkx1126

23. Ruan S, Stormo GD. Comparison of discriminative motif optimization using matrix and DNA shape-based models. BMC Bioinformatics. 2018;19: 86. doi:10.1186/s12859-018-2104-7

24. Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, Jaeger SA, et al. Diversity and complexity in DNA recognition by transcription factors. Science. 2009;324: 1720–1723. doi:10.1126/science.1162327

25. Zhao Y, Stormo GD. Quantitative analysis demonstrates most transcription factors require only simple models of specificity. Nat Biotechnol. 2011;29: 480–483. doi:10.1038/nbt.1893

26. Weirauch MT, Cote A, Norel R, Annala M, Zhao Y, Riley TR, et al. Evaluation of methods for modeling transcription factor sequence specificity. Nat Biotechnol. 2013;31: 126–134. doi:10.1038/nbt.2486

27. Cusanovich DA, Pavlovic B, Pritchard JK, Gilad Y. The functional consequences of variation in transcription factor binding. PLoS Genet. 2014;10: e1004226. doi:10.1371/journal.pgen.1004226

28. Chiu TP, Rao S, Mann RS, Honig B, Rohs R. Genome-wide prediction of minor-groove electrostatic potential enables biophysical modeling of protein-DNA binding. Nucleic Acids Res. 2017;45: 12565–12576. doi:10.1093/nar/gkx915

29. Rohs R, Jin X, West SM, Joshi R, Honig B, Mann RS. Origins of specificity in protein-DNA recognition. Annu Rev Biochem. 2010;79: 233–269. doi:10.1146/annurev-biochem-060408-091030

30. Zhou T, Yang L, Lu Y, Dror I, Dantas Machado AC, Ghane T, et al. DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. Nucleic Acids Res. 2013;41: W56–62. doi:10.1093/nar/gkt437

31. Chiu TP, Yang L, Zhou T, Main BJ, Parker SC, Nuzhdin SV, et al. GBshape: a genome browser database for DNA shape annotations. Nucleic Acids Res. 2015;43: D103–9. doi:10.1093/nar/gku977

32. Zhou T, Shen N, Yang L, Abe N, Horton J, Mann RS, et al. Quantitative modeling of transcription factor binding specificities using DNA shape. Proc Natl Acad Sci U S A. 2015;112: 4654–4659. doi:10.1073/pnas.1422023112

33. Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, Pritchard JK. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Res. 2011;21: 447–455.

doi:10.1101/gr.112623.110

34. Sherwood RI, Hashimoto T, O'Donnell CW, Lewis S, Barkal AA, van Hoff JP, et al. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. Nat Biotechnol. 2014;32: 171–178. doi:10.1038/nbt.2798

35. Qian Z, Lu L, Liu X, Cai YD, Li Y. An approach to predict transcription factor DNA binding site specificity based upon gene and transcription factor functional categorization. Bioinformatics. 2007;23: 2449–2454. doi:10.1093/bioinformatics/btm348

36. Khamis AM, Motwalli O, Oliva R, Jankovic BR, Medvedeva YA, Ashoor H, et al. A novel method for improved accuracy of transcription factor binding site prediction. Nucleic Acids Res. 2018;46: e72. doi:10.1093/nar/gky237

37. Duren Z, Chen X, Jiang R, Wang Y, Wong WH. Modeling gene regulation from paired expression and chromatin accessibility data. Proc Natl Acad Sci U S A. 2017;114: E4914–E4923. doi:10.1073/pnas.1704553114

38. Li Z, Schulz MH, Look T, Begemann M, Zenke M, Costa IG. Identification of transcription factor binding sites using ATAC-seq. Genome Biol. 2019;20: 45. doi:10.1186/s13059-019-1642-2

39. Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, et al. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. Proc Natl Acad Sci U S A. 2002;99: 757–762. doi:10.1073/pnas.231608898

40. Hemberg M, Kreiman G. Conservation of transcription factor binding events predicts gene expression across species. Nucleic Acids Res. 2011;39: 7092–7102. doi:10.1093/nar/gkr404

41. Wenger AM, Clarke SL, Guturu H, Chen J, Schaar BT, McLean CY, et al. PRISM offers a comprehensive genomic approach to transcription factor function prediction. Genome Res. 2013;23: 889–904. doi:10.1101/gr.139071.112

42. Koudritsky M, Domany E. Positional distribution of human transcription factor binding sites. Nucleic Acids Res. 2008;36: 6795–6805. doi:10.1093/nar/gkn752

43. Haynes BC, Maier EJ, Kramer MH, Wang PI, Brown H, Brent MR. Mapping functional transcription factor networks from gene expression data. Genome Res. 2013;23: 1319–1328. doi:10.1101/gr.150904.112

44. Ma S, Snyder M, Dinesh-Kumar SP. Discovery of Novel Human Gene Regulatory Modules from Gene Co-expression and Promoter Motif Analysis. Sci Rep. 2017;7: 5557. doi:10.1038/s41598-017-05705-2

45. G. TEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45: 580–585. doi:10.1038/ng.2653

46. Chen J, Rozowsky J, Galeev TR, Harmanci A, Kitchen R, Bedford J, et al. A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals. Nat Commun. 2016;7: 11101. doi:10.1038/ncomms11101

47. Shi W, Fornes O, Mathelier A, Wasserman WW. Evaluating the impact of single nucleotide variants on transcription factor binding. Nucleic Acids Res. 2016;44: 10106–10116. doi:10.1093/nar/gkw691

48. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012;337: 1190–1195. doi:10.1126/science.1222794

49. Davie K, Jacobs J, Atkins M, Potier D, Christiaens V, Halder G, et al. Discovery of transcription factors and regulatory regions driving in vivo tumor development by ATAC-seq and FAIRE-seq open chromatin profiling. PLoS Genet. 2015;11: e1004994. doi:10.1371/journal.pgen.1004994

50. Wang J, Zhuang J, Iyer S, Lin XY, Greven MC, Kim BH, et al. Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. Nucleic Acids Res. 2013;41: D171–6. doi:10.1093/nar/gks1221

51. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S, et al. AmiGO: online access to ontology and annotation data. Bioinformatics. 2009;25: 288–289. doi:10.1093/bioinformatics/btn615

52. Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Res. 2019;47: D419–D426. doi:10.1093/nar/gky1038

53. Schmiedel BJ, Singh D, Madrigal A, Valdovino-Gonzalez AG, White BM, Zapardiel-Gonzalo J, et al. Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. Cell. 2018;175: 1701–1715.e16. doi:10.1016/j.cell.2018.10.022

54. Ding SL, Royall JJ, Sunkin SM, Ng L, Facer BA, Lesnar P, et al. Comprehensive cellular-resolution atlas of the adult human brain. J Comp Neurol. 2016;524: 3127–3481. doi:10.1002/cne.24080

55. Hodge RD, Bakken TE, Miller JA, Smith KA, Barkan ER, Graybuck LT, et al. Conserved cell types with divergent features in human versus mouse cortex. Nature. 2019;573: 61–68. doi:10.1038/s41586-019-1506-7

56. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016;44: W90–7. doi:10.1093/nar/gkw377

57. Doan RN, Bae BI, Cubelos B, Chang C, Hossain AA, Al-Saad S, et al. Mutations in Human Accelerated Regions Disrupt Cognition and Social Behavior. Cell. 2016;167: 341–354 e12. doi:10.1016/j.cell.2016.08.071

58. de Mendoza A, Sebe-Pedros A. Origin and evolution of eukaryotic transcription factors. Curr Opin Genet Dev. 2019;58-59: 25–32. doi:10.1016/j.gde.2019.07.010

59. Koonin EV. Viruses and mobile elements as drivers of evolutionary transitions. Philos Trans R Soc Lond B Biol Sci. 2016;371. doi:10.1098/rstb.2015.0442

60. Catarino B, Hetherington AJ, Emms DM, Kelly S, Dolan L. The Stepwise Increase in the Number of Transcription Factor Families in the Precambrian Predated the Diversification of Plants On Land. Mol Biol Evol. 2016;33: 2815–2819. doi:10.1093/molbev/msw155

61. Larroux C, Luke GN, Koopman P, Rokhsar DS, Shimeld SM, Degnan BM. Genesis and expansion of metazoan transcription factor gene classes. Mol Biol Evol. 2008;25: 980–996. doi:10.1093/molbev/msn047

62. Paps J, Holland PWH. Reconstruction of the ancestral metazoan genome reveals an increase in genomic novelty. Nat Commun. 2018;9: 1730. doi:10.1038/s41467-018-04136-5

63. Lynch VJ, Leclerc RD, May G, Wagner GP. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. Nat Genet. 2011;43: 1154–1159. doi:10.1038/ng.917

64. Erkenbrack EM, Maziarz JD, Griffith OW, Liang C, Chavan AR, Nnamani MC, et al. The mammalian decidual cell evolved from a cellular stress response. PLoS Biol. 2018;16: e2005594. doi:10.1371/journal.pbio.2005594

65. Liu X, Somel M, Tang L, Yan Z, Jiang X, Guo S, et al. Extension of cortical synaptic development distinguishes humans from chimpanzees and macaques. Genome Res. 2012;22: 611–622. doi:10.1101/gr.127324.111

66. Sousa AMM, Meyer KA, Santpere G, Gulden FO, Sestan N. Evolution of the Human Nervous System Function, Structure, and Development. Cell. 2017;170: 226–247. doi:10.1016/j.cell.2017.06.036

67. Somel M, Franz H, Yan Z, Lorenc A, Guo S, Giger T, et al. Transcriptional neoteny in the human brain. Proc Natl Acad Sci U S A. 2009;106: 5743–5748. doi:10.1073/pnas.0900544106

68. Somel M, Liu X, Khaitovich P. Human brain evolution: transcripts, metabolites and their regulators. Nat Rev Neurosci. 2013;14: 112–127. doi:10.1038/nrn3372

69. Miller DJ, Duka T, Stimpson CD, Schapiro SJ, Baze WB, McArthur MJ, et al. Prolonged myelination in human neocortical evolution. Proc Natl Acad Sci U S A. 2012;109: 16480–16485. doi:10.1073/pnas.1117943109

70. Lacruz RS, Bromage TG, O'Higgins P, Arsuaga JL, Stringer C, Godinho RM, et al. Ontogeny of the maxilla in Neanderthals and their ancestors. Nat Commun. 2015;6: 8996. doi:10.1038/ncomms9996

71. Smith TM, Tafforeau P, Reid DJ, Pouech J, Lazzari V, Zermeno JP, et al. Dental evidence for ontogenetic differences between modern humans and Neanderthals. Proc Natl Acad Sci U S A. 2010;107: 20923–20928. doi:10.1073/pnas.1010906107

72. Gunz P, Neubauer S, Maureille B, Hublin JJ. Brain development after birth differs between Neanderthals and modern humans. Curr Biol. 2010;20: R921–2. doi:10.1016/j.cub.2010.10.018

73. Neubauer S, Hublin JJ, Gunz P. The evolution of modern human brain shape. Sci Adv. 2018;4: eaao5961. doi:10.1126/sciadv.aao5961

74. Banerjee-Basu S, Packer A. SFARI Gene: an evolving database for the autism research community. Dis Model Mech. 2010;3: 133–135. doi:10.1242/dmm.005439

75. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019;47: D1005–D1012. doi:10.1093/nar/gky1120

76. Trubetskoy V, Pardiñas AF, Qi T, Panagiotaropoulou G, Awasthi S, Bigdeli TB, et al. Mapping genomic loci implicates genes and synaptic biology in schizophrenia. Nature. 2022;604: 502–508. doi:10.1038/s41586-022-04434-5

77. Satterstrom FK, Kosmicki JA, Wang J, Breen MS, De Rubeis S, An J-Y, et al. Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. Cell. 2020;180: 568–584.e23. doi:10.1016/j.cell.2019.12.036

78. Paciorkowski AR, Traylor RN, Rosenfeld JA, Hoover JM, Harris CJ, Winter S, et al. MEF2C Haploinsufficiency features consistent hyperkinesis, variable epilepsy, and has a role in dorsal and ventral neuronal developmental pathways. Neurogenetics. 2013;14: 99–111. doi:10.1007/s10048-013-0356-y

79. Zhou WZ, Zhang J, Li Z, Lin X, Li J, Wang S, et al. Targeted resequencing of 358 candidate genes for autism spectrum disorder in a Chinese cohort reveals diagnostic potential and genotype-phenotype correlations. Hum Mutat. 2019;40: 801–815. doi:10.1002/humu.23724

80. Nowakowska BA, Obersztyn E, Szymanska K, Bekiesinska-Figatowska M, Xia Z, Ricks CB, et al. Severe mental retardation, seizures, and hypotonia due to deletions of MEF2C. Am J Med Genet B Neuropsychiatr Genet. 2010;153B: 1042–1051. doi:10.1002/ajmg.b.31071

81. Schizophrenia Working Group of the Psychiatric Genomics, Consortium. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014;511: 421–427. doi:10.1038/nature13595

82. Mitchell AC, Javidfar B, Pothula V, Ibi D, Shen EY, Peter CJ, et al. MEF2C transcription factor is associated with the genetic and epigenetic risk architecture of schizophrenia and improves cognition in mice. Mol Psychiatry. 2018;23: 123–132. doi:10.1038/mp.2016.254

83. Lai CS, Fisher SE, Hurst JA, Vargha-Khadem F, Monaco AP. A forkhead-domain gene is mutated in a severe speech and language disorder. Nature. 2001;413: 519–523. doi:10.1038/35097076

84. Chen YC, Kuo HY, Bornschein U, Takahashi H, Chen SY, Lu KM, et al. Foxp2 controls synaptic wiring of corticostriatal circuits and vocal communication by opposing Mef2c. Nat Neurosci. 2016;19: 1513–1522. doi:10.1038/nn.4380

85. Maricic T, Gunther V, Georgiev O, Gehre S, Curlin M, Schreiweis C, et al. A recent evolutionary change affects a regulatory element in the human FOXP2 gene. Mol Biol Evol. 2013;30: 844–852. doi:10.1093/molbev/mss271

86. Atkinson EG, Audesse AJ, Palacios JA, Bobo DM, Webb AE, Ramachandran S, et al. No Evidence for Recent Selection at FOXP2 among Diverse Human Populations. Cell. 2018;174: 1424–1435 e15. doi:10.1016/j.cell.2018.06.048

87. Sissler M, González-Serrano LE, Westhof E. Recent Advances in Mitochondrial Aminoacyl-tRNA Synthetases and Disease. Trends Mol Med. 2017;23: 693–708. doi:10.1016/j.molmed.2017.06.002

88. Fine AS, Nemeth CL, Kaufman ML, Fatemi A. Mitochondrial aminoacyl-tRNA synthetase disorders: an emerging group of developmental disorders of myelination. J Neurodev Disord. 2019;11: 29. doi:10.1186/s11689-019-9292-y

89. Guo M, Yang X-L, Schimmel P. New functions of aminoacyl-tRNA synthetases beyond translation. Nat Rev Mol Cell Biol. 2010;11: 668–674. doi:10.1038/nrm2956

90. Barcia G, Rio M, Assouline Z, Zangarelli C, Roux C-J, de Lonlay P, et al. Novel FARS2 variants in patients with early onset encephalopathy with or without epilepsy associated with long survival. Eur J Hum Genet. 2021;29: 533–538. doi:10.1038/s41431-020-00757-x

91. Duan J, Sanders AR, Moy W, Drigalenko EI, Brown EC, Freda J, et al. Transcriptome outlier analysis implicates schizophrenia susceptibility genes and enriches putatively functional rare genetic variants. Hum Mol Genet. 2015;24: 4674–4685. doi:10.1093/hmg/ddv199

92. Shi Y, Ghosh MC, Tong W-H, Rouault TA. Human ISD11 is essential for both iron-sulfur cluster assembly and maintenance of normal cellular iron homeostasis. Hum Mol Genet. 2009;18: 3014–3025. doi:10.1093/hmg/ddp239

93. Schmucker S, Martelli A, Colin F, Page A, Wattenhofer-Donzé M, Reutenauer L, et al. Mammalian frataxin: an essential function for cellular viability through an interaction with a preformed

ISCU/NFS1/ISD11 iron-sulfur assembly complex. PLoS One. 2011;6: e16199. doi:10.1371/journal.pone.0016199

94. Mayr JA, Haack TB, Freisinger P, Karall D, Makowski C, Koch J, et al. Spectrum of combined respiratory chain defects. J Inherit Metab Dis. 2015;38: 629–640. doi:10.1007/s10545-015-9831-y

95. Lill R. Function and biogenesis of iron-sulphur proteins. Nature. 2009;460: 831–838. doi:10.1038/nature08301

96. Lim SC, Friemel M, Marum JE, Tucker EJ, Bruno DL, Riley LG, et al. Mutations in LYRM4, encoding iron-sulfur cluster biogenesis factor ISD11, cause deficiency of multiple respiratory chain complexes. Hum Mol Genet. 2013;22: 4460–4473. doi:10.1093/hmg/ddt295

97. Bergman O, Ben-Shachar D. Mitochondrial Oxidative Phosphorylation System (OXPHOS) Deficits in Schizophrenia: Possible Interactions with Cellular Processes. Can J Psychiatry. 2016;61: 457–469. doi:10.1177/0706743716648290

98. Jablensky A, Angelicheva D, Donohoe GJ, Cruickshank M, Azmanov DN, Morris DW, et al. Promoter polymorphisms in two overlapping 6p25 genes implicate mitochondrial proteins in cognitive deficit in schizophrenia. Mol Psychiatry. 2012;17: 1328–1339. doi:10.1038/mp.2011.129

99. Bozza M, Bernardini L, Novelli A, Brovedani P, Moretti E, Canapicchi R, et al. 6p25 interstitial deletion in two dizygotic twins with gyral pattern anomaly and speech and language disorder. Eur J Paediatr Neurol. 2013;17: 225–231. doi:10.1016/j.ejpn.2012.09.008

100. Stepanova V, Moczulska KE, Vacano GN, Kurochkin I, Ju X, Riesenberg S, et al. Reduced purine biosynthesis in humans after their divergence from Neandertals. Elife. 2021;10. doi:10.7554/eLife.58741

101. Hill GE. Mitonuclear Compensatory Coevolution. Trends Genet. 2020;36: 403–414. doi:10.1016/j.tig.2020.03.002

102. Hill GE. Mitonuclear coevolution as the genesis of speciation and the mitochondrial DNA barcode gap. Ecol Evol. 2016;6: 5831–5842. doi:10.1002/ece3.2338

103. Tobler M, Barts N, Greenway R. Mitochondria and the Origin of Species: Bridging Genetic and Ecological Perspectives on Speciation Processes. Integr Comp Biol. 2019;59: 900–911. doi:10.1093/icb/icz025

104. Soto I, Couvillion M, Hansen KG, McShane E, Moran JC, Barrientos A, et al. Balanced mitochondrial and cytosolic translatomes underlie the biogenesis of human respiratory complexes. Genome Biol. 2022;23: 170. doi:10.1186/s13059-022-02732-9

105. Couvillion MT, Soto IC, Shipkovenska G, Churchman LS. Synchronized mitochondrial and cytosolic translation programs. Nature. 2016;533: 499–503. doi:10.1038/nature18015

106. Mendez FL, Poznik GD, Castellano S, Bustamante CD. The Divergence of Neandertal and Modern Human Y Chromosomes. Am J Hum Genet. 2016;98: 728–734. doi:10.1016/j.ajhg.2016.02.023

107. Endicott P, Ho SYW, Stringer C. Using genetic evidence to evaluate four palaeoanthropological hypotheses for the timing of Neanderthal and modern human origins. J Hum Evol. 2010;59: 87–95. doi:10.1016/j.jhevol.2010.04.005

108. Gómez-Robles A. Dental evolutionary rates and its implications for the Neanderthal-modern human

divergence. Sci Adv. 2019;5: eaaw1268. doi:10.1126/sciadv.aaw1268

109. Zeidan J, Fombonne E, Scorah J, Ibrahim A, Durkin MS, Saxena S, et al. Global prevalence of autism: A systematic review update. Autism Res. 2022;15: 778–790. doi:10.1002/aur.2696

110. Charlson FJ, Ferrari AJ, Santomauro DF, Diminic S, Stockings E, Scott JG, et al. Global Epidemiology and Burden of Schizophrenia: Findings From the Global Burden of Disease Study 2016. Schizophr Bull. 2018;44: 1195–1203. doi:10.1093/schbul/sby058

111. Linscott RJ, van Os J. An updated and conservative systematic review and meta-analysis of epidemiological evidence on psychotic experiences in children and adults: on the pathway from proneness to persistence to dimensional expression across mental disorders. Psychol Med. 2013;43: 1133–1149. doi:10.1017/S0033291712001626

112. Yevshin I, Sharipov R, Valeev T, Kel A, Kolpakov F. GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments. Nucleic Acids Res. 2017;45: D61–D67. doi:10.1093/nar/gkw951

113. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17: 261–272. doi:10.1038/s41592-019-0686-2

114. Hunter JD. Matplotlib: A 2D Graphics Environment. Computing in Science Engineering. 2007;9: 90–95. doi:10.1109/MCSE.2007.55

115. Waskom M, Gelbart M, Botvinnik O, Ostblom J, Hobson P, Lukauskas S, et al. mwaskom/seaborn: v0.11.1 (December 2020). 2020. doi:10.5281/zenodo.4379347

116. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018;19: 15. doi:10.1186/s13059-017-1382-0

117. Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. Mol Syst Biol. 2019;15: e8746. doi:10.15252/msb.20188746

118. Chippada B. forceatlas2: Fastest Gephi's ForceAtlas2 graph layout algorithm implemented for Python and NetworkX. Github; Available: https://github.com/bhargavchippada/forceatlas2

119. Jacomy M, Venturini T, Heymann S, Bastian M. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. PLoS One. 2014;9: e98679. doi:10.1371/journal.pone.0098679

120. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics. 2013;14: 128. doi:10.1186/1471-2105-14-128

121. Pinero J, Bravo A, Queralt-Rosinach N, Gutierrez-Sacristan A, Deu-Pons J, Centeno E, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res. 2017;45: D833–D839. doi:10.1093/nar/gkw943

122. Nishida K, Frith MC, Nakai K. Pseudocounts for transcription factor binding sites. Nucleic Acids Res. 2009;37: 939–944. doi:10.1093/nar/gkn1019

123. Yamagishi ME, Shimabukuro AI. Nucleotide frequencies in human genome and fibonacci numbers. Bull Math Biol. 2008;70: 643–653. doi:10.1007/s11538-007-9261-6

124. Fantom Consortium, the, Riken Pmi, Clst, Forrest AR, Kawaji H, Rehli M, et al. A promoter-level mammalian expression atlas. Nature. 2014;507: 462–470. doi:10.1038/nature13182

125. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods. 2013;10: 1213–1218. doi:10.1038/nmeth.2688

126. Liu S, Zibetti C, Wan J, Wang G, Blackshaw S, Qian J. Assessing the model transferability for prediction of transcription factor binding sites based on chromatin accessibility. BMC Bioinformatics. 2017;18: 355. doi:10.1186/s12859-017-1769-7

127. Cohen NM, Kenigsberg E, Tanay A. Primate CpG islands are maintained by heterogeneous evolutionary regimes involving minimal selection. Cell. 2011;145: 773–786. doi:10.1016/j.cell.2011.04.024

128. Long HK, Sims D, Heger A, Blackledge NP, Kutter C, Wright ML, et al. Epigenetic conservation at gene regulatory elements revealed by non-methylated DNA profiling in seven vertebrates. Elife. 2013;2: e00348. doi:10.7554/eLife.00348

129. Diehl AG, Boyle AP. Conserved and species-specific transcription factor co-binding patterns drive divergent gene regulation in human and mouse. Nucleic Acids Res. 2018;46: 1878–1894. doi:10.1093/nar/gky018

130. Shibata Y, Sheffield NC, Fedrigo O, Babbitt CC, Wortham M, Tewari AK, et al. Extensive evolutionary changes in regulatory element activity during human origins are associated with altered gene expression and positive selection. PLoS Genet. 2012;8: e1002789. doi:10.1371/journal.pgen.1002789

131. Paten B, Herrero J, Beal K, Fitzgerald S, Birney E. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. Genome Res. 2008;18: 1814–1828. doi:10.1101/gr.076554.108

132. Cooper GM, Stone EA, Asimenos G, Nisc Comparative Sequencing Program, Green ED, Batzoglou S, et al. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. 2005;15: 901–913. doi:10.1101/gr.3577405

133. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9: R137. doi:10.1186/gb-2008-9-9-r137

134. Blanco E, Farré D, Albà MM, Messeguer X, Guigó R. ABS: a database of Annotated regulatory Binding Sites from orthologous promoters. Nucleic Acids Res. 2006;34: D63–7. doi:10.1093/nar/gkj116

135. Griffith OL, Montgomery SB, Bernier B, Chu B, Kasaian K, Aerts S, et al. ORegAnno: an open-access community-driven resource for regulatory annotation. Nucleic Acids Res. 2008;36: D107–13. doi:10.1093/nar/gkm967

136. Portales-Casamar E, Swanson DJ, Liu L, de Leeuw CN, Banks KG, Ho Sui SJ, et al. A regulatory toolbox of MiniPromoters to drive selective expression in the brain. Proc Natl Acad Sci U S A. 2010;107: 16589–16594. doi:10.1073/pnas.1009158107

137. Portales-Casamar E, Kirov S, Lim J, Lithwick S, Swanson MI, Ticoll A, et al. PAZAR: a framework for collection and dissemination of cis-regulatory sequence annotation. Genome Biol. 2007;8: R207. doi:10.1186/gb-2007-8-10-r207

138. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine

Learning in Python. J Mach Learn Res. 2011;12: 2825–2830. Available: https://www.jmlr.org/papers/v12/pedregosa11a.html